DEVELOPMENT OF TOOLS

FOR THE ANALYSIS OF MESSAGES

IN CONTROLLED SOCIAL NETWORK ENVIRONMENTS

by

William C. Garrard

B.S. Information Sciences and Technology, Pennsylvania State University, 2009

M.S. Information Technology, Rensselaer Polytechnic Institute, 2010

Submitted to the Graduate Faculty of

The School of Information Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2017

UNIVERSITY OF PITTSBURGH

SCHOOL OF INFORMATION SCIENCES

This dissertation was presented

by

William C. Garrard

It was defended on

April 27, 2017

and approved by

Hassan Karimi, Ph.D., Professor, School of Information Sciences

Stephen Hirtle, Ph.D., Professor, School of Information Sciences

Armando Rotondi, Ph.D., Associate Professor, Center for Behavioral Health and Smart Technology, University of Pittsburgh

Dissertation Advisor: Michael Spring, Ph.D., Associate Professor, School of Information Sciences

ABSTRACT

DEVELOPMENT OF TOOLS

FOR THE ANALYSIS OF MESSAGES

IN CONTROLLED SOCIAL NETWORK ENVIRONMENTS

William C. Garrard

University of Pittsburgh, 2017

There is sometimes more demand for the attention of healthcare providers than there is supply to go around. This study evaluates a way to make expert mental health social workers more efficient at the task of moderating controlled access social network discussion boards. Sometimes, moderators need to make authoritative posts on these boards known as interventions. These are useful when needed but unnecessary interventions may degrade the benefits of organic discussion. For this study an automated decision aiding system (ADAS) tool was developed which provided the automated analysis and visualization of messages and message sentiment. This tool was designed as a means to make the expert moderators more efficient so more individuals could utilize a discussion board without proportional increase in expert moderators and the associated expense. This study determined that the custom designed automated decision-aiding system had no significant effect on participants determining if messages from such a discussion board are deserving of an intervention response for the measures of accuracy, elapsed time, or judgement confidence. The abstraction of context provided by the ADAS in this study is suspected to explain the lack of significant results, and future work would focus on identifying the level of context supply humans would require for the ADAS to have an effect.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to first thank my research advisor and committee chair Dr. Michael Spring for his help and guidance over the years. I have learned a lot from our work together and without his support this dissertation would not have been possible.

I would also like to thank my dissertation committee members Dr. Karimi, Dr. Hirtle, and Dr. Rotondi, whose time and feedback throughout this process has been greatly appreciated.

Furthermore I want to extend thanks to the faculty of the School of Social Work, the School of Health and Rehabilitation Sciences, and the School of Nursing who let me into their classes to recruit for this study, especially Dr. Mary Rauktis, without whom this would not have been possible.

I want to acknowledge the help provided by Brittney Neely and Carolyn Lamorte, without whose expertise this dissertation would not have been possible.

I want to thank my friends and family for their support, especially game night folks, Chris Lauver, and my mother Kathleen Kane.

Finally I want to give special and eternal thanks to my wife Beth, who has been an unending source of support, encouragement, and inspiration on a constant basis throughout.

CHAPTER 1: INTRODUCTION

1.1 PROBLEM STATEMENT AND CONTRIBUTION

In the healthcare system, the role of service providers can include physicians, nurses, social workers, and so on while the service seekers can include patients, their relatives, and their dependents. A single individual capable of providing a service can only provide a certain quality of service to some maximum limit of service seekers. This limit is defined by the nature of the services involved, how the parties are able to communicate, and other practical considerations. While a physical therapist could only physically interact with so many people in a day, a social worker might be able to correspond electronically with many more in the same period of time.

In this work, we examined ways in which care providers can increase the number of individuals they are able to interact with and provide care for, in the same amount of time, with the aid of automated tools. Specifically, we examined the automated analysis and visualization of messages and message author behavior on controlled access online social network discussion forums designed to provide cognitive behavioral therapy. This study sought to develop tools which can be used to identify the need for interventions. Historically, this work is performed by moderators who manually interact with users. Moderators utilize their own judgement and experience to analyze user behavior and decide when to interact with users to maintain a constructive social environment.

A web-based visualization tool was developed that experimental participants were able to use in conjunction with their own judgement and experience to identify and display user behavior for analysis and reference supplementing the need for detailed manual evaluation. The tool was based on established natural language processing (NLP) techniques. This could reduce the amount of time a moderator need dedicate to any one user in such a setting, and should increase the user load a single moderator can handle in a given amount of time.

The tool was used with experimental participants to view messages collected in a previous study which is described in section 3.1. The tool identified the messages to present based on a number of factors and the participants chose whether each message was worthy of an intervention. The participants rated their confidence in the tool's ability to provide useful information during their task and their confidence in their judgements while using the tool.

1.2 LIMITATIONS AND DELIMITATIONS

- Limitation: The data set used in this study came from a study which targeted individuals with schizophrenia or schizoaffective disorder. The techniques used may not be applicable to individuals with other disorders.

- Delimitation: The data used in this study come from individuals screened and granted access to a controlled website providing educational materials and a moderated social network.

- Limitation: The findings may not generalize to open websites and social networks.

1.3 DEFINITION OF TERMS

1.3.1 Communication

The Oxford English Dictionary defines communication as "The transmission or exchange of information, knowledge, or ideas, by means of speech, writing, mechanical or electronic media, etc." [1]. In this study, communication occurs between users. On the online social network platforms used in this study, the medium for communication is written English text. An exchange is communication occurring between two or more individuals. An exchange may be unidirectional or multidirectional, meaning users may or may not communicate back and forth to each other. Additionally, communications in exchanges may or may not be addressed to any particular user or group of users. For the purposes of this study, a dialogue is a bidirectional communication between exactly two users on a social network. Dialogues are identified when a user replies to a communication made by another user. The actual content of the communication is of no consequence, i.e. a user need not specifically identify that they are addressing or replying to a given user. Rather, the position of the user's communication in a forum thread structure as a child of a previous communication groups both the child and parent into a dialogue. Therefore, one communication can be involved in multiple dialogues.

1.3.2 Social Network Site

A social network site is defined by Boyd as “web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system.”[2]. Users on social network sites generally have some ability to make and view connections with other users, and to communicate with them.

1.3.3 Controlled Social Network Site

A controlled social network site is a social network site which is under the guidance of moderators who have a vested interest in the performance of the social network site to achieve some predetermined goal. Furthermore, a controlled social network site is not open to the public and users undergo some measure of vetting before being allowed access. Therefore, the users can be expected to have a vested interest in the ability of the social network site to help them achieve some predetermined goal.

1.3.4 Moderator

A moderator is a human agent tasked with overseeing the activity taking place on a controlled social network and to intervene when they deem it necessary to maintain the integrity and constructive nature of the exchanges taking place. Moderators are not users, but do interact

directly in user exchanges in order to conduct their activities. Moderators choose when and how to intervene in user exchanges on their own.

1.3.5 Intervention/Therapeutic Intervention

A therapeutic intervention in a controlled social network site setting occurs when a moderator deems it necessary to directly address a user by some means, typically a written message, for the purpose of maintaining the constructive nature of discussion, the direction of discussion, correct misinformation, or other issues that might affect the ability of users to utilize the social network site to its fullest potential. A therapeutic intervention is distinctly different from a social message, which moderators may also engage in. The difference is that a social comment does not fulfil any of the criteria of a therapeutic intervention. For example, a therapeutic intervention may be something like "Your last message was off topic; please only post something relevant to the current topic." A social comment may be something like "Sounds like you have some fun plans for the holiday, hope you have a good time." In this study, the unqualified term "intervention" is equivalent to "therapeutic intervention".

1.3.6 Natural Language

Natural language is the system of human communication which grew organically over time and which does not necessarily conform to any sort of logically designed language. The same information can be expressed in natural languages in a variety of different ways. Examples include English, French, and Chinese. In contrast, programming languages like C or FORTRAN

are not natural languages as they are tightly controlled in vocabulary and syntax and adhere to a series of constructed rules of use.

1.3.7 Activity Analysis

Activity analysis is the study of subject activity in any situation. In the context of this study, the participants are users on websites, and so the analysis of their activity occurs after they have utilized a website. Data involved in the analysis are usually collected automatically by the system hosting the website, such as web logs that provide timestamps on resource requests for each user, as well as content added to the system by the users manually, like forum posts.

1.3.8 Forum

The Oxford English Dictionary defines a forum as “place of public discussion”[3]. On a website, a forum enables the asynchronous text-based communication between many users. Users are able to start their own conversations on any topic or participate in one already created by responding to other users. One conversation on a forum is called a ‘thread’ and creating or adding to a thread is called ‘posting’.

1.3.9 Weblog

A weblog (or log) is a file that is stored on a web server onto which information about the server activity is recorded automatically and in real time. This activities being recorded can be the result of scheduled automatic tasks or user activity. For example, a server records in a weblog

each time a user clicks on a link, sending a resource to their browser. Weblogs can contain a variety of information and are highly customizable. Typical data recorded include datetime stamps, IP addresses, requested resources, and usernames.

1.3.10 AJAX

AJAX (Asynchronous JavaScript + XML) is a series of technologies working in parallel to provide a more interactive online experience. AJAX can be used to make web pages respond to user input quickly; pages do not need to reload entirely for a new piece of information to be displayed. AJAX is widely utilized by high profile web applications like Google Maps [4].

1.3.11 DSW

The Daily Support Website (DSW) was constructed for a study conducted by University of Pittsburgh researchers to understand the impact of a centralized web-based resource system for people with schizophrenia and schizoaffective disorder (the clients) and their families. The DSW hosted a controlled discussion forum where the clients were able to communicate with each other. This study utilizes the messages from that discussion forum as the primary data source. The design and construction of the DSW predates and is not part of this study. Furthermore, this study is not an assessment of the DSW in any way.

CHAPTER 2: LITERATURE REVIEW

This study combines techniques from several areas. Background on tools, concepts, and techniques in these areas is provided in this chapter. Section 2.1 covers Website Activity Analysis, section 2.2 covers Natural Language Processing, section 2.3 covers Studies of Online Behavior, section 2.4 covers Visualization of Social Discussions, section 2.5 covers Automated Decision-Aiding Systems, and section 2.6 covers Expert Decision Making.

2.1 WEBSITE ACTIVITY ANALYSIS

An overarching theme of this study is the way in which the activity of users on a website can be analyzed. User website activity can be recorded and reconstructed from automated logging of website resource requests. This section describes the various methods used in conducting website activity analysis.

2.1.1 Activity Tracking

Activity tracking is the systematic recording of user activity on a website. User activity can be defined as the low-level or 'mechanical' interaction with the website and its functionality,

regardless of higher-level intent. This includes for example pages where a user enters and leaves a website, how much time is spent on a page, or paths traversed during a website visit [5].

One long-used method of activity tracking is transaction log analysis (TLA), where server-side logs are used to record a user's actions during a visit. Server log entries are created automatically when a user requests a resource from a server, and can include information such as the date, time, the resource being requested, how it is being requested, from where it is being requested, the browser being used, and/or the user's name or other identification [6]. These entries only represent "physical expression of communication exchanges" [6], meaning user perception, satisfaction, or frustration are not measurable by TLA alone [7]. These logs are large in volume and automated methods are practical for analysis. There are a large number of uniquely designed and setting-focused automated TLA methods in the literature [8] [9] [10] [11].

The various methods found in the literature all, at some level, go through what Jansen [6] describes as the three phases of TLA: collection, preparation, and analysis. Research questions very often indicate what information is necessary to be collected by the server e.g. whether or not to obtain a user name for each entry, or whether knowing which page a user is coming from is necessary. In preparation, flat-file logs are cleaned (removing unnecessary data), and entered into a more useful environment, like a relational database.

In analysis, this database is queried to obtain useful information. Again, Jansen breaks this into three levels: term, query, and session analysis. In term analysis, log entries are queried for given terms. Valuable results from this include high-frequency terms and unique terms in the logs. If the terms chosen correspond to known page names, this can show which pages are being seen most often, for example. Query analysis relates to search activity where a user has entered some search terms into an interface. Here useful results can be the most frequent query terms,

unique queries, and query complexity. In session analysis, the logs are used to identify individual user sessions, this occurs when a user accessed the website, utilized it for some time, then stopped. By assigning an arbitrary time threshold between entries originating from the same source or user, which is defined as indicating a cessation of activity, those entries can be grouped into sessions. User sessions can give details like the average time users spend on the site, how many requests are made during sessions, and how often users return to the site [6].

In more modern websites which use dynamic or on-demand interactive features such as JavaScript or AJAX (Asynchronous JavaScript and XML), conventional TLA may not tell the whole story [12]. An additional level of detail can be obtained by using AJAX to report events to the same or another server. This is the basis of Google Analytics which embeds JavaScript on pages a user wants information about. Taken to the extreme, this can include all the different types of user input, like mouse clicks, mouse movement, keystrokes, and so on [13]. A number of academic uses of such technology have been published. UsaProxy by Atterer, et al [12] [14] sits in between the client browser and the server delivering website content and when it detects HTML content being sent to the client adds in custom JavaScript instructions which facilitate activity tracking. Activity recorded by UsaProxy includes mouse position, click, and hover events, page load and resize events, page focus, blur, and unload events, scroll bar use events, and keyboard input, all on the client side. Collected information is reported back to UsaProxy for storage in a database. In [12] and [14], Atterer reports that UsaProxy performed well as an unobtrusive solution to tracking user activity on websites utilizing Ajax and JavaScript, and user activity could accurately be reconstructed from their collected log data.

In [15], Kiciman and Livshits introduce another solution involving an HTTP proxy called AjaxScope. This system is designed to introduce custom JavaScript for the purpose of providing

an end-to-end view of website performance to developers for the purpose of debugging. They also posit that AjaxScope could be used in production environments for continuous improvement of website experiences by passing a small percentage of traffic through on a constant basis, rather than only being used in individual usability studies.

2.2 NATURAL LANGUAGE PROCESSING

This study utilized Natural Language Processing (NLP) tools created for sentiment analysis. NLP is a broad field that addresses many issues related to the automated processing of natural human language. In this section, background on these kinds of analysis, the tasks associated with NLP, and the tools this study used are described.

2.2.1 Sentiment Analysis

The goal of sentiment analysis is to automatically extract from natural language the overall sentiment (emotional direction or feeling) of a word, phrase, sentence, or combination thereof. This emotional direction or "sentiment polarity" is expressed in terms of a group of text being considered positive, negative, or neutral (or objective) [16]. Sentiment analysis identifies whether the author is expressing a positive or negative sentiment. Sentiment analysis is of particular value when there are a large number of simple messages on a topic. For example, amateur reviews of movies products, music, etc. which might be too numerous to examine and quantize manually.

SentiWordNet, developed by Baccianella, Esuli, and Sebastiani [17], is the result of the automatic classification of synonym sets (synsets) from the lexical resource WordNet. Synonym sets are groupings of words of similar meanings (synonyms) which convey equivalent sentiment when used in natural language. An automated algorithm annotates these synsets and produces a ready-referenceable list of positive, negative, and objective values for individual word usages. A word may have different meanings in verb and noun forms and so each form would have a different set of values. These values are decimals which when added come to equal 1. Therefore, when analyzing a set of text, the values of each word in that text from SentiWordNet can be extracted and combined, for example averaged, and then presented as the overall sentiment polarity of the target text [17].

The approach of cleaning the input of sentiment classification systems of non-useful words and sentences is explored by Pang and Lee [18]. Input is first classified as subjective or objective on a sentence by sentence basis, and then only the subjective portions are used as input in the sentiment classification system. This approach improved the observed accuracy of sentiment classification and also reduced the amount of processing needed to produce a result.

This same overall approach was used by Wilson, et al. [19]. The system identified phrases which contained “subjectivity clues” or words that usually indicate subjectivity in a phrase or sentence rather than objectivity. Again, objective phrases are considered to not contribute to the overall sentiment polarity of the text. With the resultant extract of subjective phrases and sentences, the sentiment polarity is computed with the effects of context preserved. The effect of subjective words in phrases being negated, intensified, or modified by other words in those phrases is preserved. This was shown to yield a more accurate result of sentiment

classification than baseline approaches of only computing sentiment classification on a-priori polarity values of subjective words.

Nasukawa examined the case of single phrases containing multiple diametric sentiments [20]. The method involves the dissolution of these sorts of phrases (e.g. “The product is good but overpriced”). They are extracted, broken apart, and then analyzed separately to produce a sentiment classification more indicative of the intention of the text than would otherwise be obtained by taking the text blindly as a whole.

Sentiment analysis has been used in recent years to analyze the content of social media messages (tweets, postings on forums, Facebook, etc.) as well as controlled publications like news and edited blogs in order to gain a look at the public disposition toward a variety of subjects [16] [21] [22] [23]. A marketing firm may for example utilize sentiment analysis on tweets relating to a product being promoted to identify what the public loves or hates about either the product or the promotional campaign so that future iterations of either can have that feedback involved without the costly and perhaps inaccurate process of engaging small numbers of product users directly for their opinions.

When using real-world data sources for this sort of analysis, accuracy of automated sentiment classification can be extremely low. If an organization conducting this analysis has the manpower, it might prefer more costly but reliable manual analysis of public sentiment [22]. One can therefore imagine that data from a more controlled environment, especially where users are interacting in earnest self-interest, would yield more accurate automated results from the same classification methods.

2.2.2 Semantria

Semantria is the primary NLP suite that was used in this work. Created by Lexalytics, Semantria is a popular and well known sentiment NLP tool suite and processes "billions of unstructured documents, every day, globally" [24]. It offers sentiment analysis, entity extraction, categorization, and clustering. It is offered on a trial basis for free and integrates easily with MS Excel. Semantria takes in free form text and outputs a numerical sentiment score typically between -2 (negative sentiment) and +2 (positive sentiment) with 0 representing neutral sentiment [25]. There is much work utilizing Semantria for sentiment analysis [26][27][28][29][30][31][32] or evaluating the usefulness of Semantria against other sentiment analysis tools as well as assessing the validity of its output [33][34][35] as discussed below.

In [33], Semantria is assessed against three other tools: TheySay, WEKA, and Google Prediction. The four tools were also split into commercial (Semantria, TheySay) and non-commercial (WEKA, Google Prediction) groups and compared on input from online healthcare surveys. While the non-commercial group was judged to perform better than the commercial group, Semantria was identified as being ideal for business use and commended for its ease of use. In [34], Semantria is compared to one other sentiment mining tool, Social Mention, based on input from three major social network platforms: Friendfeed, Twitter, and Facebook. Semantria was found to me more neutral in its sentiment judgements than Social Mention.

In [35], Semantria is evaluated along with four other sentiment mining tools, Text2Data, Meaningcloud, Sentirate, and Umigon. It was given Twitter data as input and was judged to perform well. It was again observed that Semantria had the tendency to judge most input as neutral while other tools went toward a mixture of neutral/positive or neutral/negative. A key

aspect of the Semantria working algorithm was also supported: decomposing complex input, judging sentiment, then recombining for an overall score gives generally better results [35]. Overall, Semantria is borne out by the literature to be a popular, easy to use, conservative, and reliable tool for NLP and sentiment extraction from informal short text.

This study utilized the free version of Semantria available on their website which integrates with Microsoft Excel. In the free version, the maximum size of an input string is 2048 bytes. In the dataset, there were 8 messages which exceeded this limit. It was observed by manual experimentation that when these messages were split in two such that the two sub-messages were under the input size limit, the output of the two parts were approximately equal in all cases. Therefore, the average of the two scores was inserted into the database for those messages.

2.3 STUDIES OF ONLINE BEHAVIOR

An overarching theme of this study is improving the way in which the moderators of social networks are able to perform their tasks. To this end, the following sections describe various studies of online behavior. Background is given on studies of methods by which online environments are kept positive and collaborative, as an important assumption of this study is the data used is coming from such an environment. Also presented are studies of methods for measuring “usefulness” of online forums.

2.3.1 Maintaining Constructive Environments in Collaborative Settings

Online discussion forums are by their nature collaborative and are often designed to be highly constructive environments, meaning contributors can gain in some way from the experience of interacting on them [36] [37]. There are many situations in which the two-way asynchronous communication afforded by online discussion forums is beneficial. For example instructors and students discussing in depth and specific class issues on their own time that would otherwise take an inordinate amount of valuable class-time to cover. Forums also allow communication to be preserved and accessed by future users who may not have been present or interested in the discussion when it first occurred [36].

Central to the efficient working of collaborative and constructive online environments is trust [38]. This is especially true in large scale free-access settings like Wikipedia, where lack of trust would cause the need for prohibitively restrictive bureaucratic constructs [38]. The freedom afforded to contributors in environments where contributors act in good faith is greatly responsible for attracting more and better contributors and contributions. Developers have created other more tangible and decentralized measures to ensure this trust, including voting systems that promote only quality material and systemic recognition of good contributions [38]. Importantly, these systems are designed to enforce trust not by the hand of administrators but rather the contributors and users [38].

Systems like these can be so enticing to the body of contributors that attempts to change them from can be met with significant resistance from the contributor “establishment”. In the case of Wikipedia, good-faith contributors and users attempting to combat the growing subversive and/or malicious “vandals” fought at length over modifications to editorial policies

[39]. "Egalitarian" users cling to the open policies that have been a hallmark of the site since its founding while others demanded the creation of privileged super-reviewers, creating hierarchy where there had been none before and potentially disrupting the highly democratic appeal of the site and driving away contributors who did not want to deal with red-tape in a volunteer environment [39].

The need for some measure of moderation and administration in collaborative and constructive online environments must be balanced against the potential harm that excessive or heady-handed authoritative action can cause. In [40], a study at North-West University in South Africa found that contributors to a free-speech discussion forum were repulsed by increasing interventions from moderators. The development of dissident speech was suppressed by interventions and the goal of a meaningful critical thinking and constructive dialogue from opposing viewpoints languished [40]. The trust amongst contributors and between contributors and moderators is destroyed when excessive interventions occur. However, since there will always be a need for interventions at some point, the ability of moderators to identify these situations becomes significant [37] [40].

2.3.2 Measuring Usefulness of Online Forums

The simplest and earliest methods for objectively measuring the usefulness of online forums to the contributors are predominately frequency-based analyses, while subjective analyses rely on self-reporting through surveys [41]. Methods for evaluating the usefulness of online forums based on the content contributed by the users came only later [36]. The content of online discussion forums has been described as "leaner" than real-life synchronous communication

between peers due to the cutting out of nonverbal communication which enrich natural language [42] [43]. To balance this, however, there is an increased level of exactness and clarity of thought in online forums which bridges the gap in communication [44].

In some paradigms of content analysis, individual sentences written by users are classified on their purpose or intent. In [45], Henri classified contributions into four dimensions: social, interactive, metacognitive, and cognitive. These refer to statements on oneself, someone else, how one reasons, and clarifications or judgements. In [46] and [47], Burnett created a typology to break contributions into either non-interactive or interactive. Interactive contributions are broken into non-collaborative or hostile and collaborative, and collaborative contributions are broken into announcements, queries, and requests. These sorts of classifications, when applied, can be counted for each user. Garrison, et al. posited that when users were found to have high levels of cognitive activity on an online forum, that forum was more beneficial to that user [48].

Along this same line, Gunawardena, et al. introduced the Interaction Analysis Model (IAM), a reworking of the Henri classification structure in [45] to detect the "construction of knowledge" [49]. Their classifications relate to what they called "phases of knowledge construction": sharing and comparing information, exploration of dissonance, negotiation of meaning, testing and modification, and phrasing agreements and applying new knowledge. Thus classified, the distribution of these phases in the content of an online forum can provide the frequency of the different phase activities [36] [49].

Newman, et al. [50] created a different system that relies on a complex series of codification of contributions on a sentence by sentence basis. Intense manual analysis of contributions is required to use this system where each code is first applied to determine which

statements in a contribution are related to critical thinking. Then, the ratio of statements contributing to critical thinking to statements detracting from critical thinking is calculated for a variety of different discussion categories. If these ratios are positive and close to 1 then the contributions have been highly related to critical thinking and therefore the online forum environment was useful to the user [50].

2.3.3 Human Understanding of Sentiment

In his “Inquiry Concerning Human Understanding”, the philosopher David Hume states “all the materials of thinking are derived either from our outward or inward sentiment”. He goes on to assert that humans resolve all thoughts and ideas, regardless of complexity, into simple combinations of feelings or sentiments. These sentiments are then figuratively attached to every idea or notion a person encounters in life, and the strength of these sentiments help people discern truth from fiction [51].

Humans rely on quick interpretation of sentiment in many situations. Sentiment can be communicated through a variety of media as well. A person can understand nation-wide sentiment on a contentious issue without physically interacting with other people through newspaper articles just as one can “read the room” to more effectively engage people face-to-face [52] [53].

Despite the quickness with which humans are able to discern sentiment in its various forms, there is still a time cost associated with doing so. In recent years, opinion sharing on the Internet has led to a glut of information that most people would not find useful to wade through, like reading several thousand individual reviews of a restaurant. Extracting and presenting

sentiment information from these sorts of data sets can aid humans in understanding other people and making decisions in their own lives [54].

2.4 VISUALIZATION OF SOCIAL DISCUSSIONS

This study incorporates visualizations built from social discussion data for the experimental participants to utilize when making decisions. To this end, this section provides background on data visualization, focused on data coming from online social discussions. Background is presented on visualizations designed to be interactive. Further background is presented on visualizations designed for various types of social network related data, including time series, complex relationships, networks, and trending data.

2.4.1 Interactive Visualizations

In the course of the analysis of any set of data, there may be call to perform automated analyses and to present the data or analysis results in a meaningful way to humans. Well-made visualizations of data allow humans to discover relationships and gain insight into data [55][56][57][58].

Visualizations allow an end user to explore the complex intrinsic relationships in data on their own terms [58][59][60]. An interactive visualization differs from other visualizations in that the user can modify the parameters being presented on demand, as opposed to a static visualization prepared from a pre-selected set of parameters [59]. In addition, an interactive

visualization may allow a user to manipulate the presented visualization itself, e.g. rotating, translating, or scaling a 3-D model. Manipulation also involves the ability of the user to remove portions of the visualization on demand, especially in the examination of 3-D models behind their façades [59]. Last, especially in simulations, an interactive visualization can allow the user to modify the data being operated upon on demand. This is useful in answering "what-if?" type questions [59].

In [55], Brunker, et al. introduce an interactive visualization tool for use in social network environments that does not rely solely on presenting static connections between users in a network graph, as many visualizations do [61][62][63]. Rather, this tool presents time based data and classifications for different users in an interactive visualization interface. That is, the authors classify the users based on the types of contributions made to a social network, then developed quantitative scores of their performance, and portrayed their activity over time in the interface [55]. Figure 1 shows an example of the interactive interface.

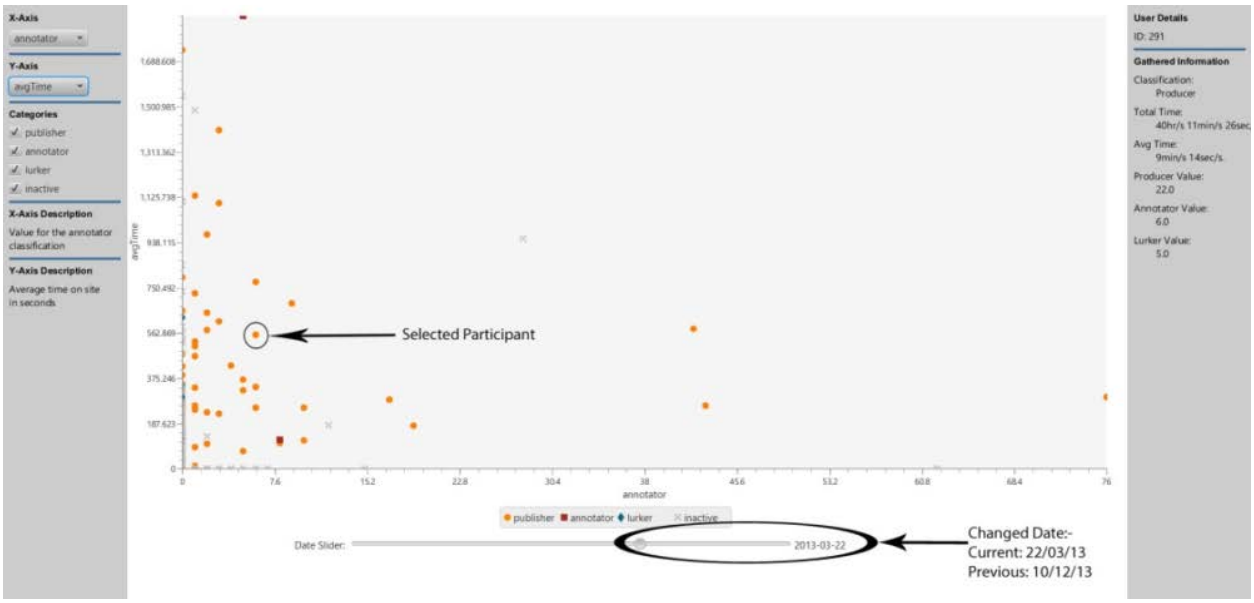


Figure 1 – An example of the interactive interface in [55]

This visualization allows a user to see multiple data points in one compact form. In the center panel of the interface is the data visualization in the form of a scatter plot. The colored dots each represent a user and the color of the dots represents the categorization of the user. Below the x-axis, in the bottom panel, there is a key for the user categorization icons and a date slider to select the point in time to be examined. Moving the date slider adjusts the data plot based on data from that date. On the left panel, x- and y-axis parameters are selected. In the right panel, further details for the user plot node selected are presented [55].

Another interactive visualization tool is presented by Tague, et al. in [64]. This tool was designed to present time series data relating to vital signs of healthcare patients. In the visualization panel, seen in Figure 2, multiple vital signs are mapped across the horizontal axis of time. The graph links are color coded to represent the deviation of vital signs from the expected

norms. Additional detail is displayed on the left hand side for selected time instances, and the whole view can be altered based on the selection of different data view ‘lenses’ [64].

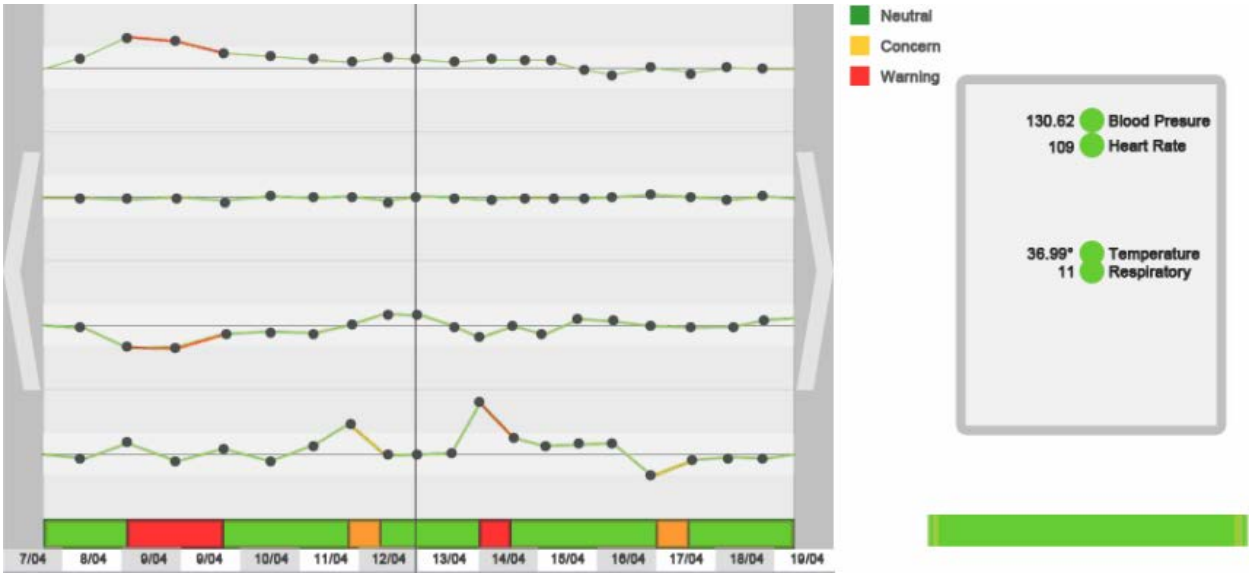


Figure 2 – An interactive visualization interface for vital signs in [64]

The visualizations in Figures 1 and 2, in addition to being interactive, focus on presenting time-oriented data. In online social networks, there is most often a time component to any sort of data, for example when exactly a message was posted on a discussion forum. Well-made visualizations for time-oriented data allows a user to examine the change of the data over time, meaning that multiple time points should be visible at once for comparison [65]. However, the time data itself is also a valid data dimension, and does not just exist to order or organize data measured in other ways [66][67].

2.4.2 Time Series Visualization

Time series data are data where a variable is sampled repeatedly over time [68]. For example, this could be the price of oil over the years, or the messages posted to a discussion forum by a user in the course of a week. There is a great volume of work published on time series visualization [68]. Time series data sets may be very large and may present meaningless visualizations unless broken down [69]. In [66], Shneiderman gives the following advice to overcome this challenge, the Visual Information Seeking Mantras: “Overview first, zoom and filter, then details-on-demand”. These aspects can be seen in Figure 3, obtained from the tool VizTree [70]. In this tool, time series data can be explored broadly and in detail.

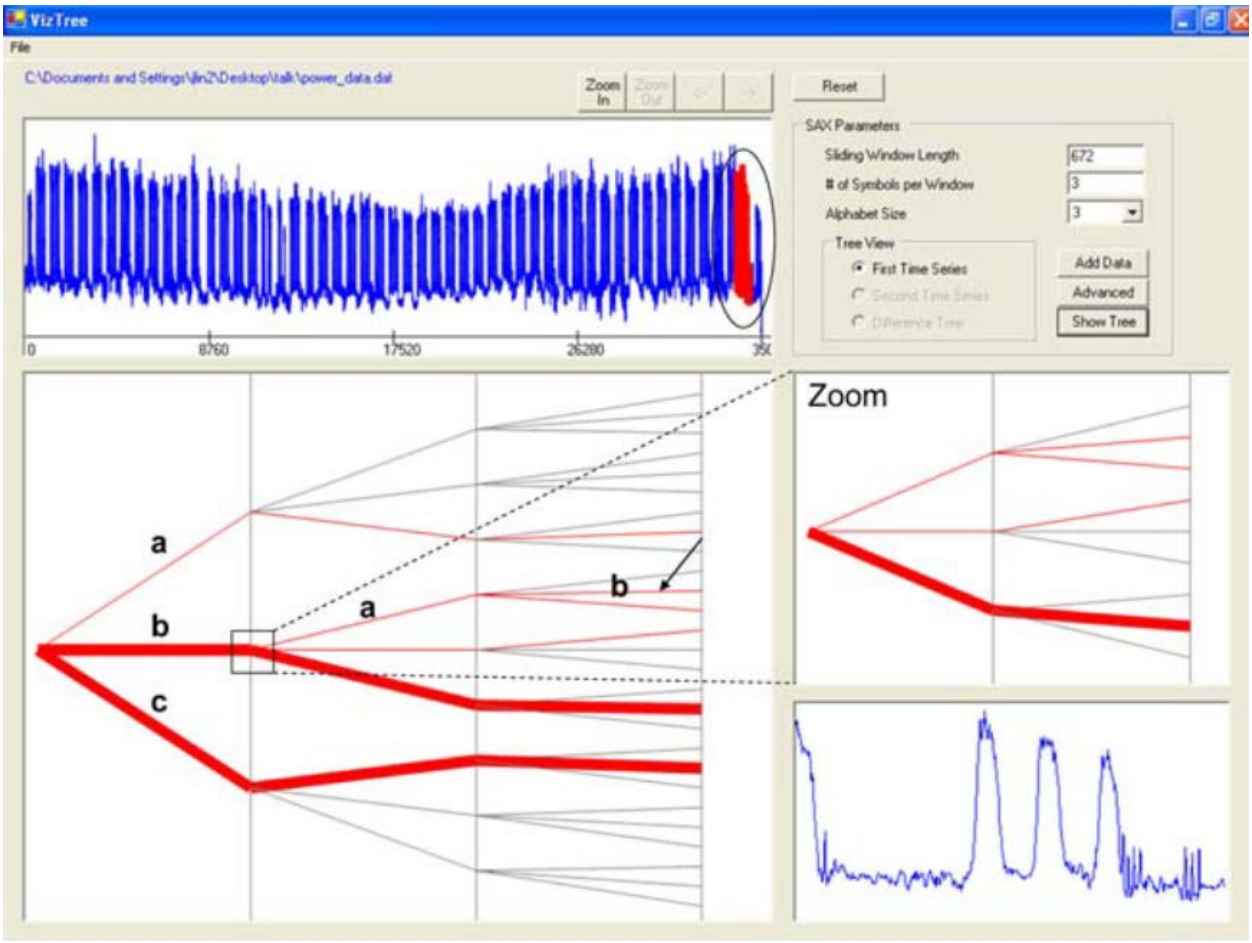


Figure 3 – VizTree tool time series visualization [70] The top panel shows the whole data series. Lower panels show details of a selected time slot.

This kind of visualization helps a user in a variety of ways. One can see coarsely where in the time series the exploration is taking place, as well as a fine view of the selected region. Additionally, the tool allows for other visualizations based on the selection to be presented in the other panels. This layout keeps all salient information on the screen at once while still being useful, in the style of a "heads-up" display.

2.5 AUTOMATED DECISION-AIDING SYSTEMS

This study involves the development of a tool designed to give experts insight into the state of clients under their guidance, and so can be described as an automated decision-aiding system (ADAS). In this section, background on the veracity and usefulness of such systems is presented. Further, the history and current state of systems of this type is discussed.

Trust is a critical component in ADAS. Experimental work has shown that humans tasked with using an ADAS usually begin with some level of trust granted to the system based on their familiarity with automated systems in general and their confidence in being able to control them. Trust was seen to decline rapidly during faulty performance, and rise back after good performance was restored, but rarely to the initial level of trust [71]. In general, humans performed better at their tasks when they felt they could trust the systems and knowledge bases they were working with than when they did not [72] [73].

As the issue of mistrust is overcome, the phenomenon of humans becoming reliant on and complacent with ADAS emerges [74]. Essentially, if an ADAS becomes sufficiently complex and sophisticated, a human will not be inclined to disbelieve its output. This condition is described as misuse of ADAS, and can lead to catastrophic failure [75]. It is recommended that humans be trained to avoid this pitfall. Experiments where humans are exposed to faulty output have been shown to sensitize them into a more realistic understanding of the capabilities of ADAS. That is, automated systems are capable of giving incorrect information without necessarily failing to work totally [74][76][77].

One of the first attempts at an ADAS in the medical field was Internist-I, developed in the 1970's at the University of Pittsburgh. Internist-I was a computer program designed to help

physicians make diagnoses in general internal medicine [78]. In contrast to other early diagnostic assisting programs, Internist-I was developed to work in medicine broadly rather than one narrow field. It operated on a large knowledge base and sophisticated heuristic algorithms to mimic the diagnostic procedures of human physicians [78]. In an experimental comparison to humans, Internist-I performed as well as clinicians, but not as well as case experts. Overall, Internist-I was judged to have impediments that led to a disrecommendation to widespread use [78].

IBM has suggested that Watson could serve as an ADAS by means of a variety of data sharing and data analytics functions in the medical field [79]. Watson has been employed by University of Texas MD Anderson Cancer Center and other medical groups in the creation of the Oncology Expert Advisor, a tool for physicians to access collected knowledge and experience to improve patient care [80][81]. Work has also proceeded on Watson's ability to aid in clinical decision making based on electronic medical records [82] [83].

2.6 EXPERT DECISION MAKING

This study addresses the challenge of helping experts to become better at making decisions. This section presents salient background work on this topic and subtopics including how experts make decisions, increasing the efficiency of expert decision making, and the use of tools by experts, as well as how these differ for non-experts.

Experts make decisions differently from novices, and this is evident in everyday life [84] [85]. Experts are relied on, or deferred to, in many different types of situations in order to put

our best foot forward. An experienced car mechanic who does not remember much of vocational school lessons may be able to diagnose an issue far faster than a recent graduate who still has lessons memorized. A chief surgeon might be brought to a patient as a consultant to identify the best course of action. Experts have the advantage of experience and intuition that novices must earn through training and/or trial and error [86].

In *Sources of Power* [86], Klein offers the following description of the importance of these advantages: "The power of intuition enables us to size up a situation quickly. The power of mental stimulation lets us understand how a course of action might be carried out. The power of metaphor lets us draw on our experience by drawing parallels between the current situation and something else we have come across."

Klein describes a paradigm of decision making, the Recognition-Primed Decision (RPD) model. There are two processes in RPD: how experts look at a situation to determine a course of action and how they evaluate their imagined course of action. Klein uses the example of firefighter activities to illustrate different ways of RPD in action. In one case, a firefighter may encounter a structure fire the likes of which has been dealt with many times before. This is called "simple match". The firefighter knows what to expect and how to handle it ("I know what is going on and what to do"). In another case, some variation or uncertainty in the situation prevents simple match, and the decision maker compares the current situation to known situations previously experienced. The reaction to the situation, once established is easy to determine ("I'm not sure what we are dealing with, but if it is X, then I know what to do"). The solution is to immediately gather more information before committing to any course of action. In a third case, it is the situation that is known, but not the reaction ("I know what we are dealing with but I don't know what to do"). Klein describes how an expert in this case may run iterative

solutions through in imagination, trying to see what might be best given what is being dealt with. Klein found RPD to be a highly relied upon decision making paradigm, especially in time critical decision making [86] [87].

Most decision-making situations have some element of uncertainty. Experts are separated from novices by being able to make good or good-enough decisions in the face of uncertainty [85]. In [88] a medical case study is presented that illustrates this difference well. A novice physician and experienced physician are both asked to review and diagnose the same patient based on the same information, recurring episodes of acute and debilitating knee pain. The novice's diagnosis is far more general than the experienced resident, citing a variety of potential maladies such as arthritis, Lyme disease, or other infection. The experienced resident is able to hone in on a single diagnosis of acute gout immediately.

This should come as no surprise but the key take away is how the experienced physician came to his conclusion. The authors claim this physician started by abstracting the discrete symptoms of the patient to see if it matched known patterns of past experiences. The physician then used this best-guess as a jumping off point to confirm with the patient other known signs of gout. The novice physician did not start with that same mental abstraction and therefore could not offer the same or any focused set of questions [88]. Both physicians may have had mastery of the same textbook knowledge of gout and other illnesses with similar symptoms. However, it is noted in [89] that "Expert clinical reasoning requires not only that pertinent clinical information be identified and meaningfully interpreted, but that it is synthesized and integrated in such a way as to enable a correct diagnosis to be made."

Decision making requires cognitive resources and as the complexity of a problem increases, the use of those resources increases as well. At some point, cognitive resources run

out and overly complex problems result in poor quality decisions [90]. To combat this, expert judgement and decision making is often aided by the use of heuristics or mental shortcuts build up from experience. Oftentimes these afford a decision maker a more efficient path to a good-enough result, but in situations where good-enough is not good enough, like medical diagnoses, these shortcuts have been criticized heavily due to the effects heuristics have on decision makers, like resistance to change or inclination to ignore "unneeded" information summarily [91][92][93][94].

CHAPTER 3: PRELIMINARY WORK

In this chapter, the work that this study is based on and the work that has been conducted to assess the efficacy of this study is discussed. Section 3.1 presents background on the Daily Support Website project which provides messages for use in this study. Section 3.2 presents exploration of the messages as a whole and section 3.3 presents exploration of intervention messages. Section 3.4 explores sentiment analysis results for intervention messages. Section 3.5 discusses trigger words in the messages. Section 3.6 discusses a heuristic for determining message history.

3.1 DAILY SUPPORT WEBSITE

The Daily Support Website (DSW) was constructed for a study conducted by University of Pittsburgh researchers to understand the impact of a centralized web-based resource system for people with schizophrenia and schizoaffective disorder (the clients) and their families. There were 199 clients in total. DSW consisted of a series of resources that the clients were able to access and use on their own time as well as a series of discussion forums where the clients could post messages to others in the study and interact with them asynchronously. There was a discussion forum for clients only, one for clients mixed with family members, and last one for

family members only. This study used the data from the discussion forum which only the clients could access. Two moderators (research staff members) were tasked with overseeing the discussion forums in order to keep the tone and tenor of the discussions progressive and constructive. The moderators each had advanced degrees and many years professional and research experience. The first moderator had master's degrees in social work and public health, was a licensed social worker, and had been a crisis counselor in outpatient mental health. The second moderator had a master's degree in social work, was a licensed social worker, and had been an outpatient mental health and addiction therapist. No one posted on discussion forums except the clients and the moderators. Figure 4 shows a screenshot of the home page for DSW. Figure 5 shows a screenshot of a piece of the discussion forum. Usernames have been censored for privacy.

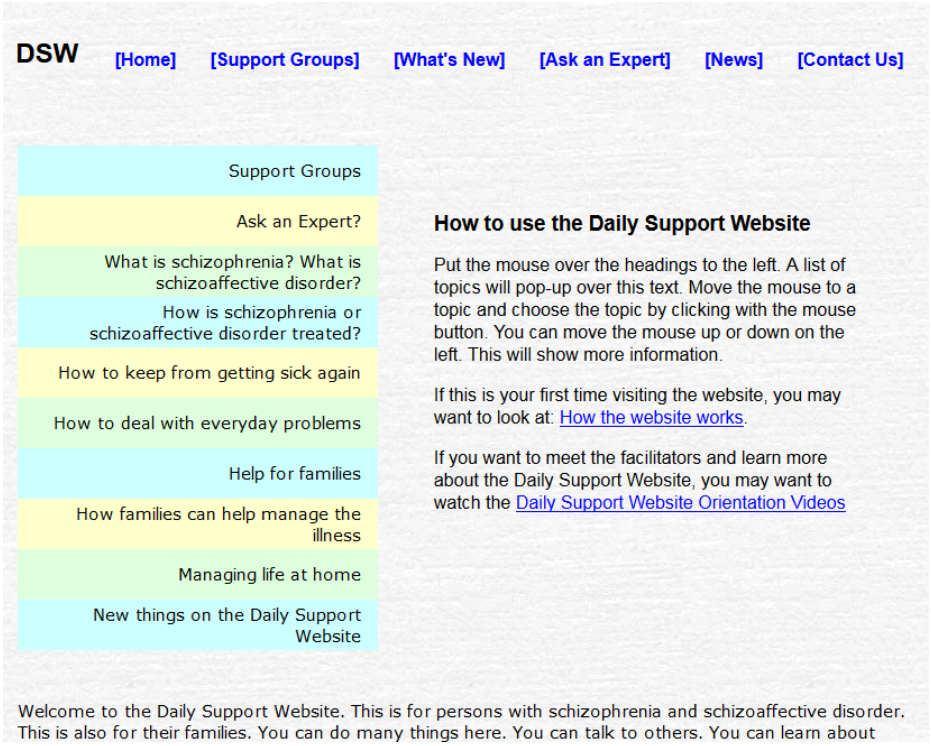


Figure 4 – DSW Home Page

Figure 5 – DSW Discussion Forum

Clients taking part in the DSW study were each assigned to case managers. There were 18 case managers in total. The case managers were social workers whose role it was to communicate directly with the clients in the study and work as an intermediary between the clients and the research staff. The case managers were responsible for helping the clients get set up with credentials to log in to the DSW and also to show them how to best use it. The case managers did not participate in the discussion forums. The discussion forum moderators were not case managers.

The DSW study also included automated alerts and questionnaires that were posted to the clients by text message, phone call, or email on a regular basis. These were set up by the case managers in a purpose built web portal separate from the main DSW site. The alerts and questionnaires were sent out on regular intervals determined by the case managers and could include medication reminders, non-medication reminders, and early warning signs questions. Early warning signs questions were designed to assess whether a client may be experiencing the early warning signs of a symptomatic episode. The answers the clients provided to these alerts and questionnaires triggered notifications to the case managers and other support contacts. Clients were given mobile phones by the study and instructed on how to use them by the case managers.

3.2 EXPLORATORY DATA ANALYSIS

This section presents low level information on the composition of the data set used in this study. This study utilized data consisting of messages from the client-only discussion forum from the DSW study. Each message in the forum is stored in a database, where each row in the database is one message. Each row has a reference to the thread topic the message was posted to as well as who the user is replying to, how many other users have replied to the message, the date and time of posting, and the message text. For this study, there is a 1.89 MB corpus of 6899 individual user messages. Tables 1-4 give detailed information on dataset statistics. Additionally, this section presents some descriptive statistics on the output from the Semantria natural language processing software. Semantria is used in this study to obtain sentiment polarity scores for each message. Semantria uses a logarithmic scale when generating sentiment polarity scores and so the full range is between $+\infty$ and $-\infty$. Most scores fall between +2 and -2. [95][96]. Table 5 gives information on the Semantria output.

Table 1 – User Statistics

Users	
Total Clients	199
Total Moderators	2
Average Topics per Client	19.52
STDEV Topics per Client	38.64
Clients Appearing in Only 1 Topic	32
Max Topics Per Client	350
Average Messages per Client	24.945
Max Messages per Client	422
STDEV Messages per Client	46.493

Table 2 – Topic Statistics

Topics	
Total Topics	1357
Average Responses Per Topic	5.089
STDEV Responses Per Topic	2.701
Topics Started by Moderators	176
Topics Started by Clients	1181
%Topics with < 10 Responses	92.99%
Max Topic Responses	20
Topics with No Response	69
% No Response Topics Started By Clients	50%
Topics with Single Client	71
Average Clients Per Topic	3.865
Max Clients Per Topic	12
STDEV Clients Per Topic	1.733

Table 3 – Time Statistics

Time	
Average Topic Time Span	5.86 days
STDEV Topic Time Span	8.46 days
Max Topic Time Span	78.27 days

Table 4 – Message Statistics

Messages	
Total Messages	6899
Total Client Messages	4988
Total Moderator Messages	1911
Total Words in Messages	267133
Average Words per Message	23
Total Sentences in Messages	21274
Average Sentences per Message	3.08

Table 5 – Semantria Statistics

Semantria	
Average Sentiment Score	0.173
Max Sentiment Score	2.136
Min Sentiment Score	-2.534
STDEV Sentiment Score	0.392
% Positive Sentiment	55.79%
% Negative Sentiment	21.67%
% Neutral Sentiment	22.52%

3.3 INTERVENTION EXPLORATORY ANALYSIS

This section provides background on the types of interventions that were carried out during the DSW study. The DSW moderators made 1911 postings to the discussion forums. These consisted of a mixture of social comments and therapeutic interventions. As defined in section 1.3.5, a therapeutic intervention and a social comment are two distinct methods of moderator interaction. The social comment does not guide or control the flow of the discussion, whereas the therapeutic intervention does. An intervention may also offer advice or give direction in direct response to the author of the message it is responding to. As such, a therapeutic intervention breaks the organic nature of the discussion while a social comment does not. An intervention is not necessarily a negative event, and it does not necessarily follow a negative message. The chief notion of an intervention is that the moderator is speaking in their capacity of authority rather than as a peer.

There is no algorithmic way to determine which of the postings made by the moderators are therapeutic interventions or social comments. All of the postings made by the moderators were analyzed manually. Direction for identifying the difference between social comment and therapeutic intervention was established over several in-person interviews with the DSW moderators. During classification two anomalies were uncovered. There were 62 instances of moderator messages being classified as interventions that were actually replies to another moderator message. Also, there were 149 instances of 0-Level topic messages (in reply to no one) posted by moderators which were classified as interventions. The moderators could have been replying to the 'gestalt' of the discussion forums with these messages, but since these messages, as well as the moderator to moderator responses do not meet the condition of being in

reply to a client, they are not be considered in this study as intervention messages. The classification results were confirmed by the moderators afterward. Overall, the DSW moderators made 342 social comments and 1358 therapeutic interventions. There were 211 messages removed from the 1911 postings for the reasons explained above.

A discrepancy relating to this total of 1358 interventions arose after data collection was complete. Unknown until that time, there were actually 66 instances of duplicate interventions. That is, a client message which was responded to by both moderators. During the initial construction of the intervention messages, these were counted twice. Therefore the true number of client messages which received interventions is 1292, but the number of moderator intervention responses is 1358. This discrepancy is illustrated in greater detail in section 5.1.4.2. Examples of the 342 social comments and 1358 therapeutic interventions are presented in Table 6 along with the messages that preceded them.

Table 6 – Examples of Therapeutic Interventions and Social Comments by Moderators in the DSW Discussion Forums

Preceding	**Therapeutic Intervention**
So, sometimes I throw myself a pity party and cry about everything that has happened this past year. I wish I didn`t hear voices. I wish the voices had not effected me the way they did. I wish I had never started medication. I wish so many things. However, what happened, happened and I have to get over myself. So, even when I`m depressed about having a mental illness I act as though I`m not. I smile most of the day. I make jokes about myself and what I`ve gone through. Pretending to be okay with everything, actually helps me almost be okay with it all. I will always wish that I didn`t have this, but I must admit that there are far worse things. How about you guys, do you get concerned over your illness too? Do you wish you didn`t have it? What makes you feel okay? What makes you forget you`ve got a mental illness?	Feelings of sadness and thinking about how things might be different are normal parts of dealing with a mental illness. But it is important to do what you are doing - identifying ways to cope and getting support.
is it weird if you keep remembering things from the past and can concentrate on the present and future?	I agree that many people spend time thinking about things from the past. Do you worry about past events, or just remember them? Some other suggestions for how to focus on the present are planning time each day to do a hobby or interest you enjoy, and making a to do list for what you want to get done each day.
how to be a friend? she told me she was at the casino with her mother in law.yet she thinks i have a problem. she asked me did i go grocery shopping? i should have mentioned i am broke. it is hard to live on a fixed income.i told her no. she had the nerve to tell me i need to apply for section 8 that pays your utilities. i really can not do her ignorance anymore.	I`m sorry to hear that you are dealing with some issues with your friend. Try not to let her suggestions upset you too much. She may be concerned for you but not showing it in the best of ways. Is this person someone who has been in your life a long time?
Preceding	**Social Comment**
hi. I`m new to the group.	Hello tee! I`m glad to see your post here. This is Carolyn, we talked on the phone last week. Did you have any difficulty getting onto the website?
Today is my birthday. A birthday is a day for celebration & reflection. While there is life there is hope for a better future & happiness. How do you celebrate your birthday?	Happy Birthday! I enjoy having cake on my birthday.

I have learned a lot on the site, and have enjoyed communicating with a lot of people and hearing about their issues. I hope to use the site some this week. It is such a shame that I missed so much time on it due to my shoulder, but I accept it. Thanks for everything. I will write some more later. Brittany and Carolyn you are great.	I am glad that you have enjoyed being a part of the website, piano. It is good to hear that your shoulder is feeling better. I hope you will be able to use the website some this week. Please let us know if you have any topics you would like to discuss before you finish up.

3.4 SENTIMENT RELATION TO INTERVENTIONS

The decision to present message sentiment in the experimental treatment case was based on the process a human uses to decipher the meaning of natural language. Humans can explicitly examine the sentiment in text, but more often it is done implicitly. Understanding sentiment is a core component of human natural language communication [51] [52] [97]. In this study, the experimental treatment case displayed information on the sentiment of the messages. Therefore, it was important to see if there is some relationship between the sentiment of messages the DSW moderators encountered and their propensity to intervene beyond that which can be assumed from common-sense.

Table 7 shows the distribution of messages that either did or did receive an intervention response grouped into five categories based on sentiment score. The groups were centered at the average observed sentiment score for the dataset and increase at ±1 standard deviation of the observed sentiment scores.

Table 7 – Intervention Distribution Grouped by Sentiment

	No Intervention	Intervention	**Total**
HighPositive	89	22	**111**
MedPositive	390	238	**628**
Neutral	2794	809	**3603**
MedNegative	338	192	**530**
HighNegative	85	31	**116**
Total	**3696**	**1292**	**4988**

Table 7 shows there is a relation between the sentiment of messages and the likelihood of an intervention taking place after. In the HighPositive category and MedPositive categories, messages were responded to with an intervention 13.71% and 14.09% of the time respectively. In the HighNegative and MedNegative categories, interventions occurred for 38.84% and 42.22% of messages respectively. Overall, it is evident that messages with more negative sentiment were responded to with interventions more often.

3.5 TRIGGER WORDS

In correspondence with the original DSW moderators, it was noted that decisions to intervene were based partially on the existence of certain themes or topics of discussion present in the messages. These themes and topics include expressions of safety issues (e.g. “I don’t care anymore, I’ll teach them, I can’t go on”), increase of symptoms (e.g. “I’m hearing more voices, I’m so anxious, I have not slept”), asking for advice (e.g. “How do I handle my anxiety at a public event?”), and asking for information (e.g. “Does anyone know if medication X has side effects”) [98][99]. The data was therefore explored to identify what sort of relationship exists

between messages with trigger words relating to these topics and messages with an intervention response. These trigger words were highlighted for participants in the treatment case.

To build a list of candidate trigger words, the literature on schizophrenia symptoms and treatment was examined. A program was developed which traversed a collection of letters from the DSW moderators and public web pages covering schizophrenia symptoms [98][99][100][101] in order to identify the most frequently used and salient words that could be considered trigger words. Definite and indefinite articles, conjunctions, prepositions, etc. were excluded. Plurals, singulars, and alternate forms of identified words were added manually to ensure meaningful matching could take place. This list of words can be found in Appendix A.1.

Table 8 shows the distribution of messages that either did or did not receive an intervention response, grouped by the prevalence of the trigger words obtained from the literature in those messages.

Table 8 – Intervention Distribution Grouped by Literature Trigger Word Presence

	No Intervention	Intervention	**Total**
Trigger word present	1445	503	**1948**
Trigger word not present	2251	789	**3040**
Total	**3696**	**1292**	**4988**

Table 8 shows that there were 2082 messages which contained at least one of the trigger words identified. Of these, 30.59% received an intervention response. There were 2906 messages that did not contain any of the identified trigger words. Of these, 22.53% received an intervention response.

In a second phase, trigger words were identified from the DSW messages directly. Messages that received an intervention response were analyzed by a word frequency program. The same cleaning function was applied to this list as to the previous. This list of words is in Appendix A.2.

Lastly, these lists were shown to the DSW moderators for their expert opinion, and it was determined that all the words identified in this preliminary work should be considered trigger words in a practical sense [102][103]. Therefore the final list of trigger words is the set union of all the words listed in Appendix A.1 and A.2.

3.6 MESSAGE HISTORY

There are some challenges related to the selection of messages to be shown to participants during the experiments. Some authors are more prolific than others, and so at a certain time point an author may have a longer posting history than other authors.

The tools being tested here would be used by moderators over a period of time in moderating a group. There would be a history that would develop for certain clients, as would be the case with or without tools. To assess the impact of the tools, messages with more history are more likely to have meaning for the moderator. At the same time, there are instances in which moderators need to make judgements when there is little history for a given client. Thus, care needs to be taken in assessing the impact of the tools based on the context within which a message occurs.

Simply selecting messages from the middle 50% or so of the time span of the data set would not be representative of reality. In practice, a moderator would not only start interacting in a discussion forum after some amount of posting has taken place. Rather, a moderator would be present from the beginning. Therefore, messages must be selected from throughout the available time span.

In order to have some measure of the degree of history the messages being displayed have compared to other users' postings, the following heuristic metric was developed for use during post experiment analysis. The degree of relative history H_r for a given message at time point T is equal to the ratio of the number of messages posted by the author M_{aut} before T to the average number of messages posted by all users M_{all} before T. $H_r = \frac{M_{aut}}{M_{all}}$.

CHAPTER 4: RESEARCH DESIGN

The design of this study is organized into two parts. They are designed to answer the research question:

> Can the automated analysis and visualization of an author's messaging behavior on a controlled access online social network discussion forum allow experts to moderate such forums more efficiently?

In the first part of the study, professional social workers were tasked with identifying the most necessary interventions they made in a discussion forum that was part of the DSW study. This serves the purpose of creating a gold-standard corpus of discussion forum messages for the study of moderator intervention decisions.

The second part of the study utilizes this corpus. Participants were drawn from the University of Pittsburgh's School of Social Work graduate program. In conversation with faculty from the School, it was determined that these graduate students can have a variety of different backgrounds and clinical experience. The students have a foundational curriculum early in the program then branch into specialized fields. Participants for this study were recruited from those students in the School who have completed the foundational curriculum. The degree of clinical experience these participants have was collected [104].

The experimental trial in the second part of this study involved dividing these participants into a control and treatment group by matched assignment. Both groups were asked to read a series of messages from the corpus created in the first part of the study. This corpus included messages posted in the discussion for which the moderators did and did not choose to intervene. The participants were tasked with making a judgement on whether the message deserves an intervention response or not. They were also asked to rate their confidence in their judgement. In the control case, the participants were only shown the message. In the treatment case, participants were shown the message and visualizations of the posting frequency of the message author, sentiment of the message history, and sentiment of the current message. The message shown also had any "trigger words" (as discussed in section 3.5) present highlighted in red.

The rest of this chapter is organized as follows. Section 4.1 describes the first part of the study and discusses the experimental workflow. Section 4.2 describes the second part of the study and discusses the experimental workflow. The study has been reviewed and approved as "Exempt" by the University of Pittsburgh Institutional Review Board, IRB# PRO16080004.

4.1 PART ONE: DSW MESSAGES

The messages come from a research study that was conducted at the University of Pittsburgh. A complete description of the experimental scenario for the study is in Section 3.1. It was designed to give individuals with symptoms of schizophrenia and schizoaffective disorder and their friends and family a centralized web-based resource portal. This includes a monitored and controlled-access discussion forum. In the forum, clients are able to post discussion topics and

talk to each other with oversight by study moderators. These moderators are able to intervene in the organic discussions in a number of ways. First, they could stimulate conversation by posting topics for discussion, including welcomes, topics of the day or week, or social topics. The moderators could also guide or shape discussions that were negative in tone, off topic, or inappropriate. The moderators were tasked with observing all content on the forums in order to identify when and where to intervene, based on their experience and judgement [105].

4.1.2 Message Preprocessing

In their original form, the messages had several features that were problematic for their use in this study. First, many messages were preserved originally with HTML tags and character entities. These HTML tags were removed programmatically. Character entities for printable characters were replaced with the plain-text they represent. Therefore, the messages presented in this study consisted only of plain-text.

The DSW messages in the database posted by test and/or dummy accounts were removed. They were created by DSW administrators and do not represent any experimental data gathered during the DSW study.

Last, these cleaned messages were given as input into the Semantria sentiment analysis tool described in section 2.2.2. The numerical sentiment score for each message was created by Semantria and then associated with the messages. This score is a distillation and abstraction of the natural language contents of a message into a single numerical value. A positive score indicates positive sentiment and a negative score indicates negative sentiment contained within the source message.

All other idiosyncrasies of the messages were left as-is, i.e. spelling errors, grammatical errors, and so on which were made by the clients when originally posting their messages were not changed or corrected in any way.

As discussed in section 3.3, responses posted by the moderators fell into two categories: social comment and therapeutic intervention. Several messages that were not responses were removed. The responses posted by the moderators were manually classified and the results of this classification were confirmed by the DSW moderators.

4.1.3 Experimental Workflow

The DSW study moderators were tasked with identifying the interventions that were most necessary. These are the same individuals who were employed by the DSW study and made these interventions originally. They were presented with the intervention responses made by both of the moderators as well as the client messages they were responding to. They were asked to rank the necessity of their intervention on a 5-point Likert scale, based on their expert opinion. It was suspected that the moderators of the DSW study may have had varying levels of certainty when deciding to intervene. The reported rankings were used to make a mixture of messages for use in the second part of the study. Selected messages are those which were ranked between three and five, inclusive.

4.1.4 Experimental Data Recorded

The experimental data recorded was the ranking for each intervention.

4.1.5 Variables and Expected Results

The independent variable was the message being presented to the DSW moderator. The dependent variable was the DSW moderator's judgement of the necessity of the intervention on a 5-point Likert scale. It was expected that not all messages would be judged by the DSW moderators to have been of highest necessity in retrospect.

4.1.6 Evaluation

An interrater reliability measure was reported for the ratings made by the DSW moderators by way of the Cohen's Kappa statistic. This statistic assumes the categories the raters are sorting the items into are ordinal. In this case, the 5-point Likert scale can be interpreted as ordinal as it represents a measure of certainty where 5>4>3>2>1. In the case of messages which were rated very differently by the DSW moderators, they were informed of the differences and asked to re-rate.

4.2 PART TWO: ASSESSING THE NEED FOR INTERVENTION

In the second part of the study the messages from part one were utilized in a decision making task. The participants were presented with the messages and asked to decide whether or not they deserve an intervention. This was meant to mimic the environment the moderators of the DSW study worked in, where they read messages from DSW users as they were written and then decided whether or not to respond. Participants in the control case only saw the messages while participants in the treatment case saw the messages as well as the visualizations of the automated decision-aiding system developed for this study.

4.2.1 Participants

Participants for the second part of the study were recruited mainly from the University of Pittsburgh's School of Social Work graduate master's program (MSW). The goal was to identify participants whose qualifications mirror those of the DSW moderators as closely as possible. Communication with faculty from the School of Social Work revealed that students in that program all go through the same foundational coursework then decide on specializations such as mental health, working with children, working with the elderly, and others. Additionally, the students had a diversity of experience working in clinical settings. Age and gender demographics of the MSW students are shown in Tables 9 and 10. Approximately one third of current MSW students have undergraduate degrees in social work while the rest have degrees in the areas of psychology, sociology, child development, criminal justice, history, or English. Approximately three quarters of MSW students are full time, with the remainder part time [104].

For this study, sixty participants who have passed the foundational phase of the MSW program were sought and separated by matched assignment into an independent treatment group and control group. Participants were recruited directly and in person from different second-year MSW courses with the permission of the course instructors.

Table 9 – School of Social Work Graduate Student Age Demographics

Age	<22	22-25	26-30	31-35	36-40	41-45	46-50	>50
Count	4	266	157	49	26	20	14	11

Table 10 – School of Social Work Graduate Student Gender Demographics

Female	464
Male	82
Unknown	1

Prior to the commencement of the experiment, the participants were given an entry questionnaire designed to measure their experience in clinical work and their familiarity with automatic decision-aiding systems. The questions in this entry questionnaire were:

1. What is your familiarity with automated decision aiding systems? (not familiar at all 1-5 very familiar)
2. How many years' experience do you have in the field of social work?
3. Are you a full time or part time social work student?
4. Are you from an undergraduate social work program or from a clinical work program background?

5. Is your field concentration direct practice or COSA?
6. Do you have personal experience with schizophrenia?
7. Have you worked in a job with exposure to schizophrenia?
8. Have you taken a class with exposure to schizophrenia?
9. How knowledgeable are you about schizophrenia (not knowledgeable at all 1-5 very knowledgeable)
10. What is your age?
11. What is your gender?

4.2.1.1 Sample Size

Justifying the size of the sample for this study is challenging due to the novelty of this research. The formula for calculating sample size in this case is $n = \left(Z_{\alpha/2} + Z_{\beta}\right)^2 * 2 * \sigma^2 / d^2$, where α is the significance level desired (often .05), β is the power level desired (often .8), Z denotes the critical values relating to those points on the Z distribution, σ is the population or sample standard deviation, and d is the estimation of the difference between the means of the groups. In practice it is often the case that historical data or educated guesses based off of past or very similar research are used for the values of the population or sample standard deviation and the difference estimation. In this study, however, no such sources exist from which to draw these values. Therefore it was recommended that the best course of action was to base the sample size for this study on other studies in the literature which evaluate the effect of automated decision-aiding systems. Most of the studies examined for this purpose utilized a sample size in the range of about 50-100 individuals (though some had more or fewer) and so the sample size of n=30 per

group for this study fits in with what other researchers in this field have done in the past. [71] [72] [73] [76] [77] [78]

4.2.2 Message Selection

The experiment in the second part of the study involved exposing each participant to 100 messages from the DSW study and asking for their judgement on whether a message deserves an intervention response. The corpus of messages created in the first part of this study was utilized as the gold-standard of messages that may or may not require intervention. Half of the messages shown to each participant were selected from the set of messages identified in the first part of the study, where an intervention occurred. The other half was messages that did not receive an intervention response. Tables 11, 12, and 13 show distributions for various message attributes. The messages were exposed to the participants in random order.

Table 11 – Client Messages with and without Intervention

Type	Frequency
Client Messages without Intervention	3696
Client Messages with Intervention	1292
Total	**4988**

Table 12 – Sentiment Score Distribution By Standard Deviation (σ = 0.38, mean = 0.13)

Standard Deviation	Frequency
<-2 σ	130
>-2 σ and <-1 σ	551
>-1 σ and < mean	2097
>mean and < 1 σ	641
>1 σ and < 2 σ	1389
>2 σ	180
Total	**4988**

Table 13 – Message Posting Timespan Distribution

Dates	Frequency
1st 5 Months	111
2nd 5 Months	457
3rd 5 Months	1840
4th 5 Months	1539
5th 5 Months	1041
Total	**4988**

4.2.3 Experimental Workflow

The experiment in the second phase of this study involved presenting each participant with 100 messages, one at a time. For each message, the participant was asked to make a judgement if the message warrants and intervention response or not. They were also asked to rate their confidence in their judgement on a 5-point Likert scale. In the control case, the participant only saw the message text. In the treatment case, the participant was presented with the visualizations

seen in Figure 10 (section 5.2) plus the message with trigger words highlighted. Figure 6 below was the original prototype treatment case interface.

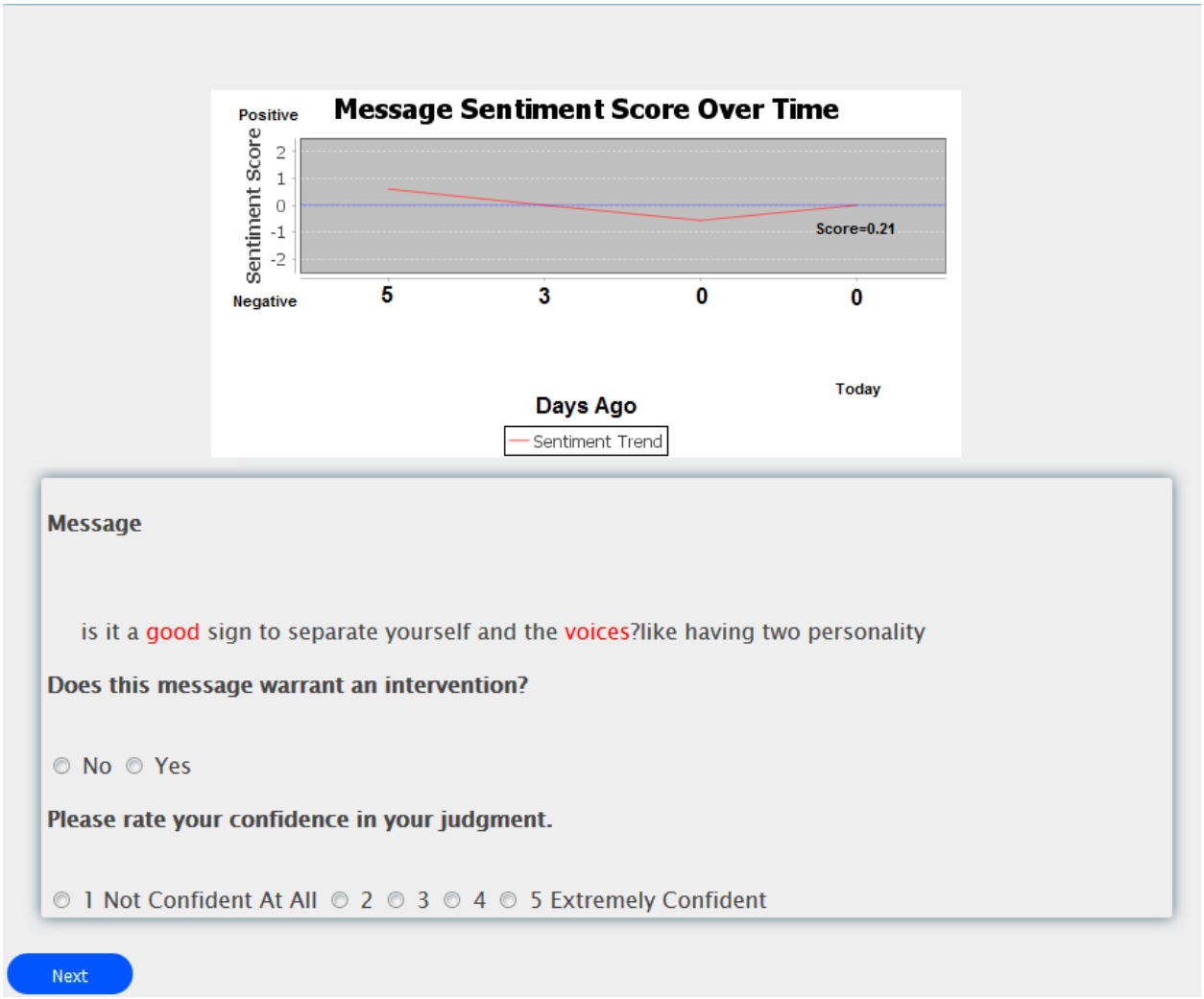


Figure 6 – Prototype Treatment Case Interface

The graphical visualization (top center) is a plot of the sentiment polarity for messages posted by the author of the current message in the past. The design is taken from similar

examples presented in [64] and [70]. The plot included at most the last ten messages posted by the author and displayed the relative time difference in days from the current message. If the author has not posted ten or more messages before the current message was posted, then all of the previous messages were plotted. The sentiment score for the current message is labeled at the last point in the graph plot. Lastly, trigger words contained within the message being displayed are highlighted in red.

After the participant has saw all the messages, they were asked to complete a short exit survey to determine their overall confidence in the visualizations' representation of reality as well as their confidence in their judgements as a whole. The questions of this exit survey were:

1. Please rate your confidence in your judgements as a whole (not confident at all 1-5 extremely confident)
2. Please rate your confidence in the visualizations as a whole (not confident at all 1-5 extremely confident) *
3. Which part of the visualizations impacted your confidence in your judgements most: the plot or the trigger word highlighting? *
4. Which part of the visualizations helped you more in making your judgement decision: the sentiment plot or the trigger word highlighting? *
5. Which kind of messages impacted your confidence in your judgements most: longer messages or shorter messages?
6. Was your confidence in your judgements impacted by having fewer than ten prior messages displayed in the plot? *
7. What kind of information would have made this task easier for you? (free response)
8. What did you like the most about the visualizations? (free response) *

9. What did you not like the most about the visualizations? (free response) *

*Treatment case only

4.2.4 Experimental Data Recorded

The experimental data that was recorded for each participant trial includes:

1. Anonymized participant id
2. Judgements (yes/no) for each message presented
3. Confidence (1-5) for each judgement made
4. Start and end time of each participant session
5. Start and end time of each message task
6. Entry questionnaire responses
7. Exit survey responses

4.2.5 Variables and Expected Results

In this experiment, the independent variable was exposure to more information about the message author in the form of the sentiment polarity trend visualization and trigger word highlighting. The dependent variables were the judgements on whether to intervene, the confidence ratings for those judgements individually, and the time taken for each message judgement.

It was expected that the participants would make more accurate intervention judgements in the treatment case and would have higher confidence in their judgements in the treatment case.

It was expected that participants would have higher confidence overall in their judgements in the treatment case. Also, it was expected that participants would spend less time making judgements in the treatment case.

4.2.6 Hypotheses

$H_{1\text{-}0}$: There is no statistically significant difference in the mean intervention judgement accuracy with respect to the gold standard made by participants between the treatment and control cases.

$H_{1\text{-}1}$: There is a statistically significant difference in the mean intervention judgement accuracy with respect to the gold standard made by participants between the treatment and control cases.

$H_{2\text{-}0}$: There is no statistically significant difference in the mean confidence ratings made by participants for each intervention judgement between the treatment and control cases.

$H_{2\text{-}1}$: There is a statistically significant difference in the mean confidence ratings made by participants for each intervention judgement between the treatment and control cases

$H_{3\text{-}0}$: There is no statistically significant difference in the mean elapsed time for each intervention judgement between the treatment and control cases.

$H_{3\text{-}1}$: There is a statistically significant difference in the mean elapsed time for each intervention judgement between the treatment and control cases.

4.2.7 Evaluation

The research goal of this study was to evaluate whether there was a difference between the treatment group and control group means for accuracy of judgements, the confidence in the judgements made, and the time taken to make the judgements. A 1-way between subjects ANOVA test was used to identify if there is a statistically significant difference in the means of the two groups. The alpha level for significance was set to p=.05. The effect size of the treatment on each of the data series was reported with partial eta squared.

CHAPTER 5: RESULTS

This chapter presents and discusses the results of this study. Section 5.1 discusses Part 1 of the study, how it was set up and conducted, and the post-hoc analysis. Section 5.2 discusses the same for Part 2 of the study as well as the formal results and discussion of the hypotheses put forth in section 4.2.6.

5.1 ASSESSMENT OF THE RESPONSES TO MESSAGES

In Part 1 of the study, the original DSW moderators (the classifiers) assessed the necessity of the intervention responses they had written. There was no experiment in this part of the study and it is not designed to establish the validity of their assessments. Rather, it was a preparatory process creating a gold standard for use in the remainder of the study. This process took approximately five weeks and was conducted remotely via a custom built website. Due to the fact that the classifiers were tasked with classifying 1358 separate messages each, the website was designed to be an 'on-demand' tool which could be accessed across multiple sessions any time either of the classifiers had free time to spend on the task. Figure 7 shows a screenshot of the interface the classifiers used to complete the task. The three buttons at the bottom of the interface allowed the classifiers to (left to right) save their current message rating and continue, to save their current

message rating and quit the current session, and to just quit the current session without saving the current message rating.

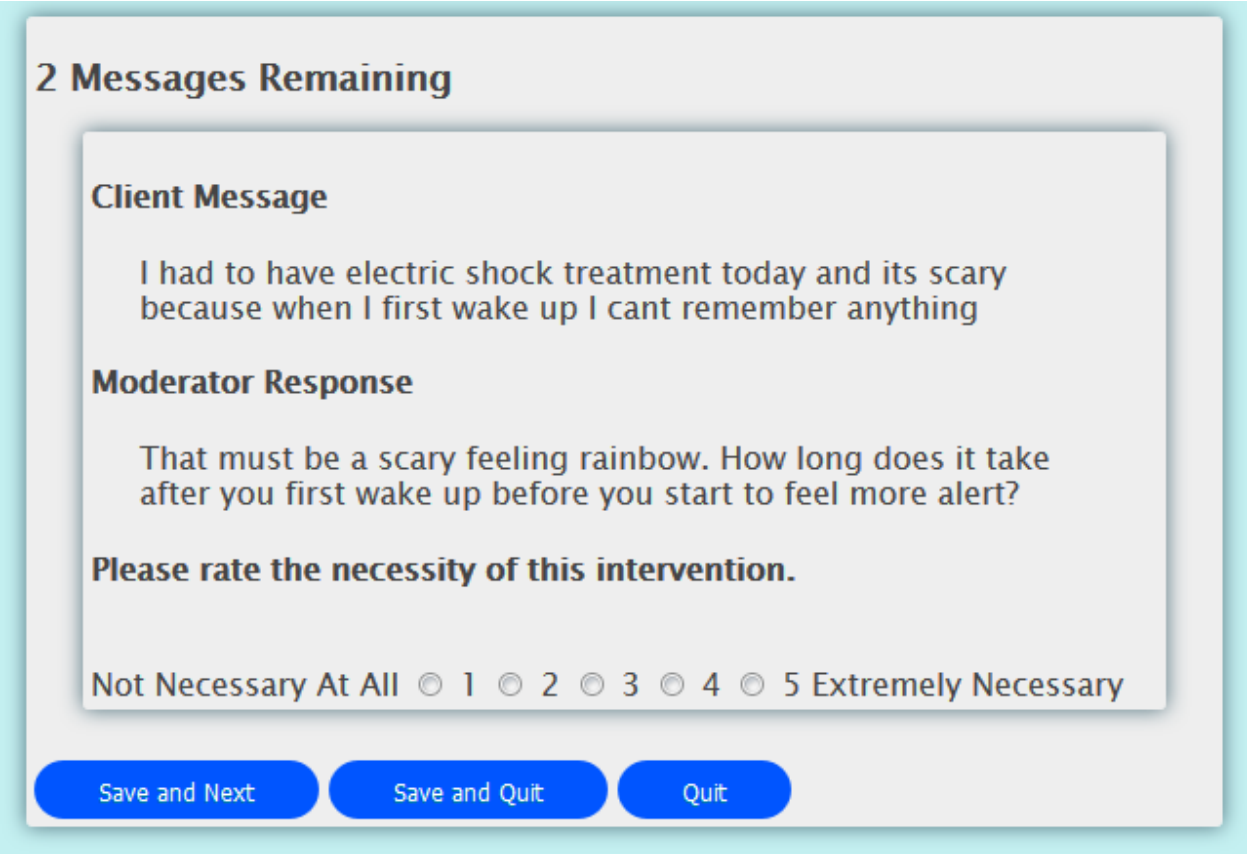


Figure 7 – Interface for Part 1 Classifiers

The classifiers began this task on November 30, 2016 and completed it on January 4, 2017, a time span of 36 calendar days. Data for classifier activity including actual elapsed time spent making the classifications, number of sessions, and classifications made per session are presented in Table 14 for each classifier.

Table 14 – Classifier Activity

Classifier	Total Sessions	Total Elapsed Time (m)	Average Session Length (m)	Average Ratings Per Session	Average Time Per Rating (m)
1	19	530.16	27.90	71.47	0.39
2	23	399.46	17.37	59.04	0.29

It had been estimated that it would take each classifier between 4-6 hours to complete the task. This was shown to be an underestimation. Classifier 1 took about 8.8 hours and classifier 2 took 6.65 hours to finish. The classifiers were compensated for the task. The distribution of ratings made by the two classifiers is shown in Figure 8.

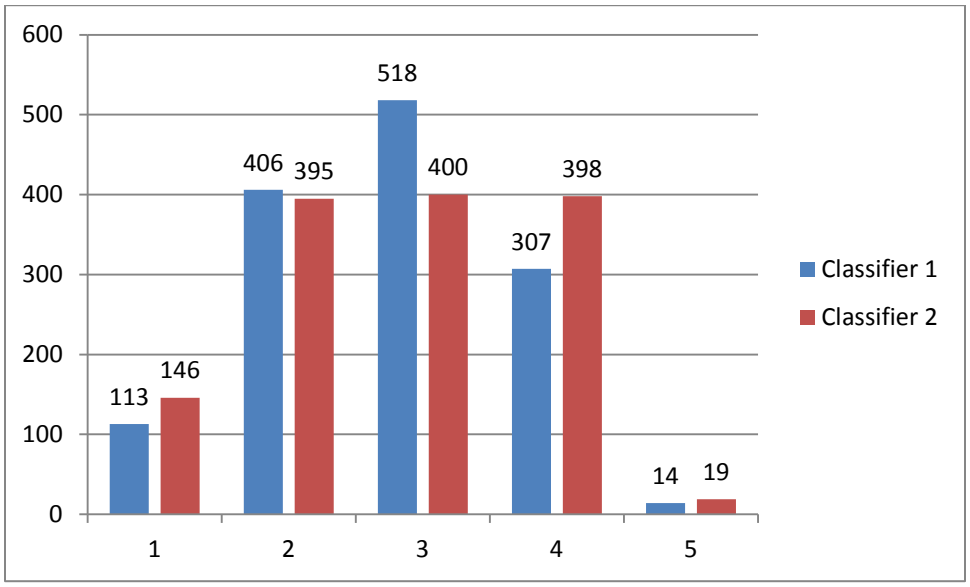


Figure 8 – Frequency of Each Rating Made by Classifier

Each classifier made the same number of ratings, 1358 in total. Most of these ratings fell in between 2-4 for both classifiers. Each had a very small number of 5 ratings compared to the total number of ratings. The average rating for classifier 1 was 2.78 and for classifier 2 was 2.81.

5.1.1 Inter-Rater Reliability

Messages that were rated as highly necessary, scores of 4 or 5, by both classifiers were examined. Table 15 shows the distribution of the ratings made for each message individually. The horizontal axis is the rating made by classifier 1 and the vertical axis is the rating made by classifier 2. Each cell contains the number of messages which received the corresponding ratings by the two classifiers. For example, the cell containing the value 119 indicates there were 119 messages which received a rating of 4 by classifier 1 and a rating of 3 by classifier 2.

Table 15 – Distribution of Ratings Made For Each Message

		Classifier 1					
	Ratings	**1**	**2**	**3**	**4**	**5**	**Total**
Classifier 2	**1**	87	54	5	0	0	146
	2	21	275	92	7	0	395
	3	5	66	325	119	0	515
	4	0	11	95	158	4	268
	5	0	0	5	20	9	34
	Total	113	406	522	304	13	1358

The diagonal-right line of cells (1,1) through (5,5) show the number of messages which both classifiers independently agreed upon for that rating. In the range that qualified for consideration for use in Part 2, there were 158 messages rated 4 and 9 messages rated 5, for a total of 167.

The total number of agreements was 854. The formula for simple inter-rater reliability (IRR) is:

$$IRR = \frac{Total\ Agreements}{Total\ Items}$$

In this case, IRR = 854/1358 = 62.9%. Generally, the literature indicates that the interpretation of simple IRR measures can be tricky [106] [107]. However, there are some benchmarking scales which have been developed for different types of IRR statistics that are flexible enough to be roughly applied to simple IRR [107]. Table 16 gives three of these: the Landis-Koch benchmark scale, the Fleiss benchmark scale, and the Altman benchmark scale.

Table 16 – Landis-Koch, Fleiss, and Altman Benchmark Scales

Landis-Koch		**Fleiss**		**Altman**	
Statistic	**Strength**	**Statistic**	**Strength**	**Statistic**	**Strength**
<0.0	Poor	<0.40	Poor	< 0.20	Poor
0.0 - 0.2	Slight	0.40 - 0.75	Intermediate to Good	0.21 - 0.40	Fair
0.21 - 0.40	Fair	> 0.75	Excellent	0.41 - 0.60	Moderate
0.41 - 0.60	Moderate			0.61 - 0.80	Good
0.61 - 0.80	Substantial			0.81 - 1.00	Very Good
0.81 - 1.00	Almost Perfect				

The main purpose of these scales is to provide some rough guide to making a subjective characterization of an objective IRR measure. In the case of this study, the score of 62.9% can be interpreted as “substantial”, “intermediate to good”, or “good” depending on the scale.

5.1.2 Cohen’s Kappa

The Cohen’s Kappa statistic is designed to provide a measure of IRR which accounts for agreement between classifiers due to chance [108]. Cohen’s Kappa is also designed for exactly 2 classifiers. The formula for Cohen’s Kappa is:

$$\kappa = \frac{P_0 - P_e}{1 - P_e}$$

Where P_0 is the observed probability of agreement and P_e is the expected probability of chance agreement. Alternatively it can be calculated from frequencies:

$$\kappa = \frac{f_0 - f_c}{N - f_c}$$

Where f_0 is the observed frequency of agreement, f_c is the calculated frequency of chance agreement, and N is the number of items which have been classified. The frequency of chance agreement is calculated by:

$$f_c = \sum_{r\,min}^{r\,max} \left[\frac{\left(\sum r_{classifier\,1} \times \sum r_{classifier\,2} \right)}{N} \right]$$

Where $r_{min}$and r_{max} are the minimum and maximum rating categories respectively, $\sum r_{classifier\,1}$and $\sum r_{classifier\,2}$are the total items rated for a single category by each classifier, and N is the total number of items which have been classified. Table 17 shows the calculation of Cohen's Kappa for this study.

Table 17 – Calculation of Cohen's Kappa

	Ratings					
	1	**2**	**3**	**4**	**5**	**Total**
Agreement (f_0)	87	275	325	158	9	**854**
By Chance (f_c)	12.14875	43.64948	56.12077	32.68336	1.397644	**146**
Cohen's Kappa	**0.584158**					

An example calculation of f_c for category 1 is as follows, and the values used can be found in Table 15:

$$\frac{(113 \times 146)}{1358} = 12.14875$$

Then, the Cohen's Kappa statistic calculation is:

$$\kappa = \frac{854 - 146}{1358 - 146} = 0.584158$$

Based on this result, the IRR measure can be categorized as “moderate” or “intermediate to good” depending on the benchmarking scale used.

5.1.3 Re-Rating

The next step in reconciling the differences in the ratings made by the classifiers was the process of re-rating. In general both classifiers should work together to come to a consensus on each item that they disagreed on. Given the intent to use only messages which had received a 4 or 5 rating by one of the classifiers, the number of messages to be re-rated was 238 rather than the 504 on which there were disagreement. More explicitly, a message which had been rated as 3 and 4 separately was included in the re-rating, but a message which had been rated as 2 and 4 separately was excluded.

The rerating process began January 31, 2017, by which time one of the classifiers was unable to participate due to personal reasons. Therefore the one classifier was tasked with re-rating all the identified messages alone. For most of the messages to be re-rated (224 of 238), the classifier changed her original ratings to match the other’s. In the remaining 14 instances, the re-rating classifier determined to change the other classifier’s original ratings. This task was conducted via a digital spreadsheet with the messages listed alongside the two ratings made, and the modified rating was filled in manually for each message by the classifier. Total time spent conducting the re-rating task was about 1 hour as reported by the classifier.

5.1.4 Message Selection

The following sections cover various issues relating to the construction and composition of the intervention and non-intervention message sets used during the data gathering phase in Part 2 of the study.

At the conclusion of the re-rating task, there were 138 responses re-rated as 4. There were 5 responses re-rated as 5. The messages corresponding to these 143 responses together with those of the 167 responses originally agreed upon as either 4 or 5 ratings produced an ideal corpus of 310 messages with intervention responses of verified high necessity to be used in Part 2 of the study. Careful examination of the data revealed duplication of some messages in the database as a result of an error when messages were moved from one database table to another. The duplicates were removed resulting in 306 intervention messages. (It is noted that the duplicate messages had been rated identically by the two classifiers.)

5.1.4.1 Non-Intervention Messages Selected

There were 3,696 messages that were rated implicitly as 0 necessity because they did not receive a moderator response. To these were added the messages from the rating task set that did not get 4/4 or 5/5 scores from the moderators. These totaled 986 again due to message duplicates. The total number of non-intervention messages was 4,682. The average rating of non-intervention messages was 1.0028 / 10.

5.1.4.2 Message Duplicates and Totals Discrepancies

The total number of messages in the intervention and non-intervention datasets was 4,988, the same total number of client messages reported in Table 4. However, a discrepancy exists as the number of messages in the Part 1 set (1,358) and the number of messages rated implicitly as 0 (3,696) sum to an incorrect 5,054. It was observed after the completion of data gathering that a number of duplicates had existed in the messages classified in Part 1.

The discrepancies described here are illustrated in Figure 9. There were 1,358 messages given to the classifiers in Part 1. However, there were only a total of 1,292 (306+986) messages used from this set. The difference is seen in 66 duplicates which existed in the Part 1 classification set. That is, these were single client messages which the moderators had responded to multiple times or both moderators had responded to individually. This discrepancy emerged due to the way the Part 1 classification set was built originally.

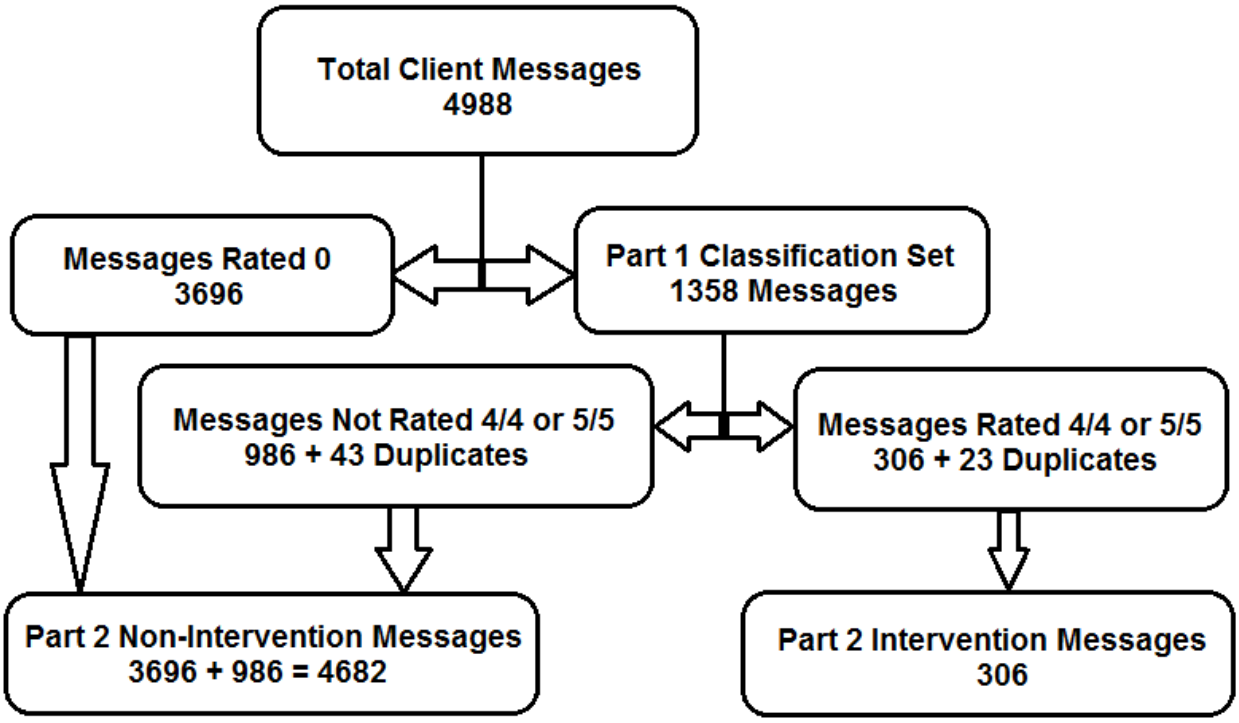


Figure 9 – Intervention/Non-Intervention Message Sets Creation Flowchart

Critically and despite these issues, there was no intermingling of message types. That is, there were no messages counted as intervention messages that should not have been and the same was true for non-intervention messages. This means the experimental design requirement of each participant being given 50 intervention and 50 non-intervention messages was preserved in all cases. Also worthy to note is the classifiers were consistent in rating duplicate messages similar to each other.

5.1.4.3 Message Rating

A rating scale was developed for the messages after data gathering was complete in order to allow for more granular analysis. The non-intervention message set contained 986 messages which were low priority intervention messages.

The rating developed for the messages was the sum of the original ratings given by the classifiers after re-rating. Thus, each message had a rating in the range of 0-10. Table 18 gives the distribution of messages by rating. There are zero messages with a rating of 9 because all intervention messages had been originally rated 4/4 or 5/5. Also there are no messages with rating of 7 because all messages in Part 1 which had been originally rated 3/4 were re-rated to either 4/4 or 3/3. Also there are no messages with rating of 1 because 0/1 was not an allowable original rating on a 5-point Likert scale.

There are different counts for messages with rating of 8. This is because intervention messages with rating of 8 came from those with an original rating of 4/4 whereas the few non-intervention messages with a rating of 8 came from those with an original rating of 3/5. Because the re-rating task only included messages with a 1 point difference, these were not included in the intervention messages.

Table 18 – Distribution of Messages by Rating

	Rating	Count
Intervention Messages	10	14
	9	0
	8	292
Non-Intervention Messages	8	3
	7	0
	6	412
	5	151
	4	265
	3	74
	2	81
	1	0
	0	3696
	Total	4988

For intervention messages, an accurate judgement would be agreeing an intervention is needed. For non-interventions rated 0, an accurate judgement would be agreeing one is not needed. The non-intervention messages with a rating between 1 and 6 are explored in more detail in section 5.2.10.

5.2 ASSESSMENT OF ADAS TOOL

In Part 2 of this study, research participants were recruited to complete a task wherein they were shown messages from the DSW discussion board and asked to judge whether or not they deserved a moderator intervention. The data gathering process consisted of 61 participant sessions and was completed in approximately 4 weeks between February 3 and March 2, 2017.

The treatment and control case interfaces used by subjects to for this task are shown in Figures 10 and 11 respectively.

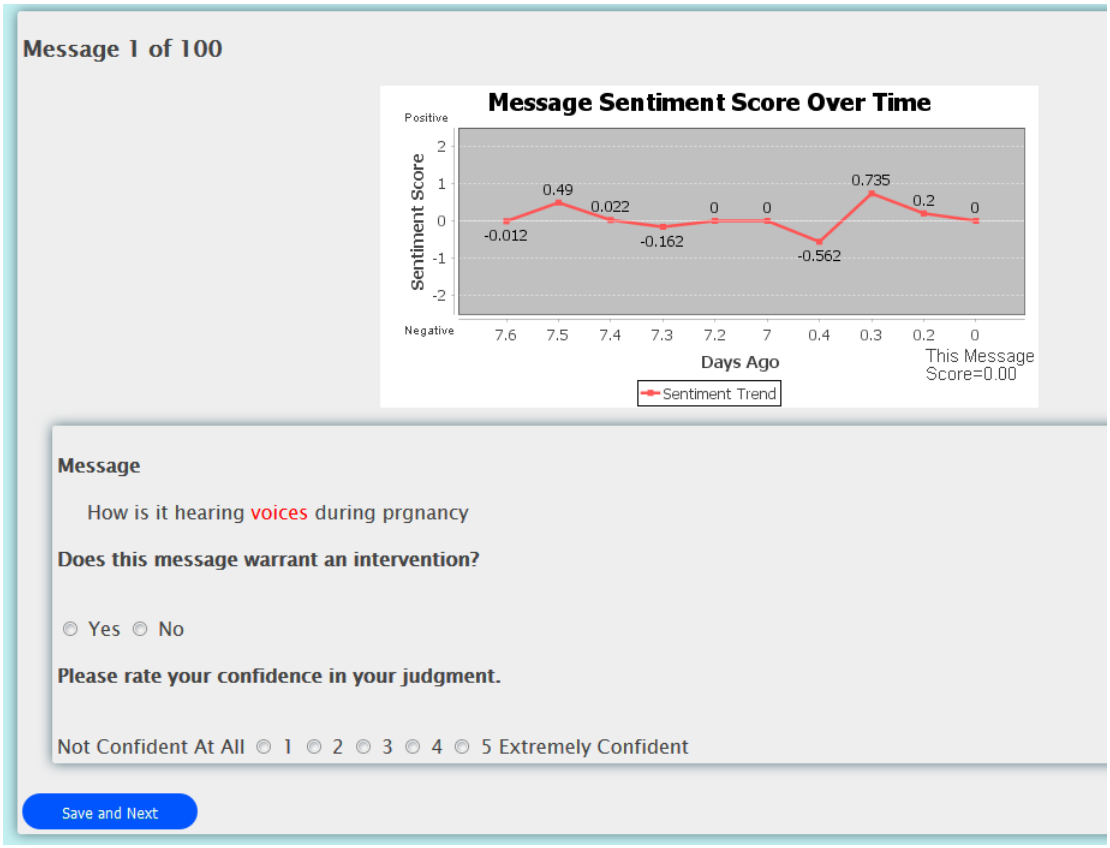


Figure 10 – Treatment Case Interface

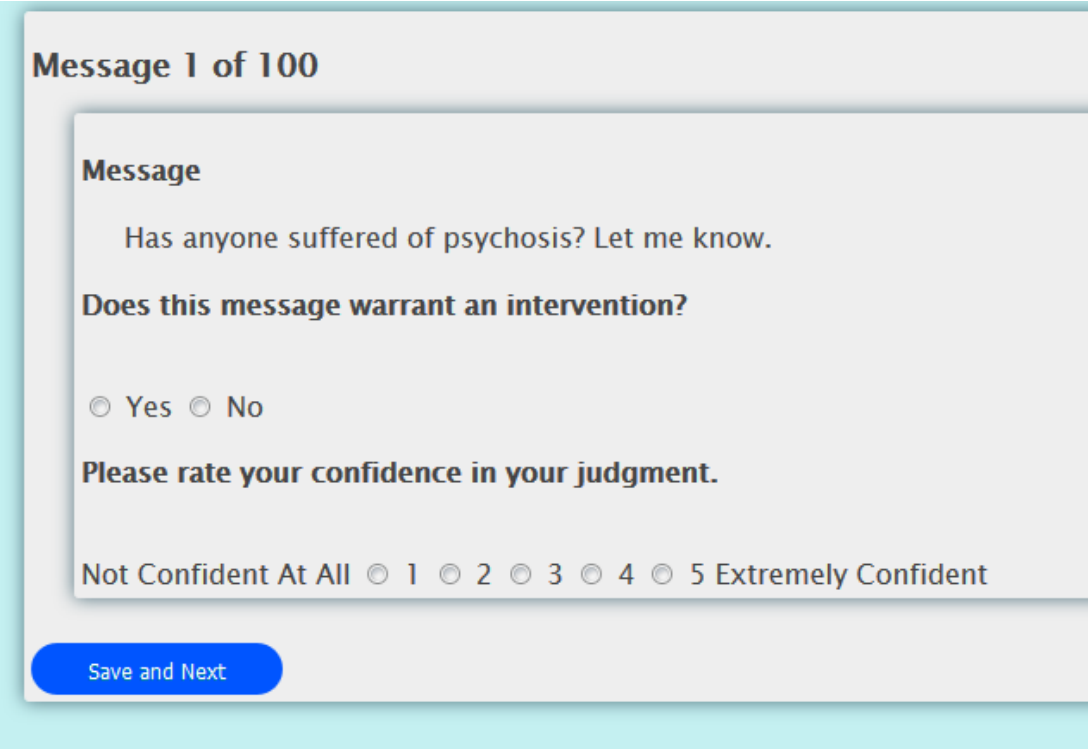


Figure 11 – Control Case Interface

The treatment case interface was refined from the original prototype interface. First, it was modified to include values for each data point labeled on the plot itself. Each data point was represented by a small square on the plot line. Also, the algorithm labelling the x-axis was written to include a dot-notation for messages occurring on the same day. For example, in Figure 10, the message being displayed is labeled as being written 0 days ago, and messages written earlier that same day are labeled 0.2, 0.3, and 0.4.

5.2.1 Participant Recruiting

The research participants were recruited from the University of Pittsburgh's School of Social Work (SSW), School of Health and Rehabilitation Sciences (SHRS), and School of Nursing (SN). Initial recruiting was conducted in-person at 10 different graduate SSW classes with the permission of the social work faculty members.

Paper handouts were distributed at each visit containing information about the study's purpose, the task the participants would be doing, time commitment, compensation, and contact information. This handout is included in Appendix B.1.

As research participants began scheduling sessions by email, reminders were included in email confirmations to organically spread the word about the study in order to attract more participants. SSW faculty members also sent email reminders to their classes at their offering and some participants took it upon themselves to advertise the study on social media.

As the number of participants who had either scheduled or completed a session reached about half the target 60, it was decided that other students with similar qualifications to graduate SSW students could be found in the Master of Counseling program at the SHRS as well as the psychiatric nursing program at the School of Nursing. A second round of in-person recruiting was conducted at two classes at the SHRS, and an email recruiting letter was sent to the small psychiatric nursing community at the School of Nursing. Table 19 summarizes the affiliation of the 61 research participants.

Table 19 – Research Participant Affiliation

Affiliation	SSW	SHRS	SN
Count	52	8	1

5.2.2 Participant Consent and Training

At the beginning of a session, the participants were given a copy of the consent script to read and keep afterward while it was read aloud to them. Per the IRB protocol for this study, signatures were not required to be obtained, and consent was obtained verbally. The consent script is included in Appendix B.2.

Participants were separated to the control or treatment group before they arrived for their session by matched assignment. For the first participant, a coin was flipped then subsequent participants were assigned to the groups alternatively based only upon when they responded with interest to participate in the study. Each participant was given training on the task they were to perform during the session. The training for the control and treatment cases was particular to the interface and two separate training documents were created for the participants to refer to during the training and throughout the session. The training documents consisted of a description of the experimental scenario and the features of the interface they were going to use. Each participant was instructed to read the training document while it was spoken aloud to them. The training documents are included in Appendix B.3 and B.4. The 61st participant was sorted into the treatment group by luck of the draw and so the treatment group has one more participant than the control group.

5.2.3 Message Truncation

During the development of this study, the database tables containing the messages from the DSW studies were copied several times for different data preparation processes. Before Part 1 of this study began as messages were transferred between tables, messages which were longer than 255 characters were truncated to that character limit. This error was not recognized until the middle of the participant sessions of Part 2 when a participant raised the issue of sentences ending abruptly. It was confirmed shortly after that all processes described in this study after the sentiment score calculations had been performed on the truncated messages. That is, the sentiment scores were based on full length messages, but the classification of Part 1 and the judgements of Part 2 were performed on the set in which some messages had been partially truncated. Approximately 42% of the intervention messages and 21% of the non-intervention messages were truncated. During the course of the study, there were 1252 (41% overall) truncated intervention messages presented during sessions and 650 (21% overall) truncated non-intervention messages presented during sessions. The average rating for truncated messages after Part 1 was 6.1 and the average rating for complete messages was 5.28.

In light of this oversight, the sections presenting the results of the three hypotheses of this study include separate analyses of messages which were whole and which were truncated.

5.2.4 Entry Questionnaire

The participants were given the entry questionnaire described in section 4.2.1 after the training was complete. Figures 12-22 show the results.

The participants were overwhelmingly female. This matches the observed gender distribution of SSW and SHRS graduate students overall. Most participants had less than 2 years of experience in their field of study ($\sigma = 1.85$, mean = 2.25) and were full time students. This was as expected given the recruiting target of graduate students.

Average age was 25 years with a standard deviation of 2.8 years. Most (n=39) were between 22 and 25 years of age. About half the participants had been exposed to schizophrenia professionally. A majority of participants had exposure to schizophrenia in an academic setting and a majority had no exposure to schizophrenia in personal life.

Despite the youthfulness of many of the participants as compared to professional social workers or counselors, the entry questionnaire showed that the participants by and large had some prior experience with or knowledge of schizophrenia. This differentiates these participants from the general population to some degree, which is in line with the research goal of having participants whose qualifications approach the qualifications of experts.

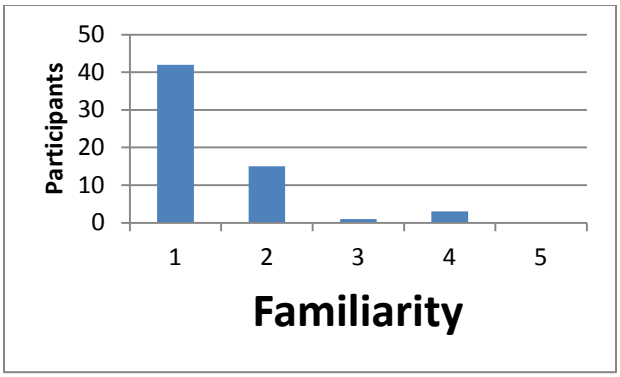


Figure 12 – Participant Familiarity with ADAS Distribution

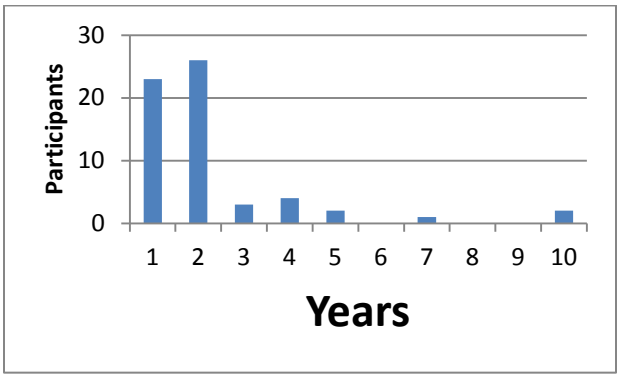


Figure 13 – Participant Years' Experience in Field of Study Distribution

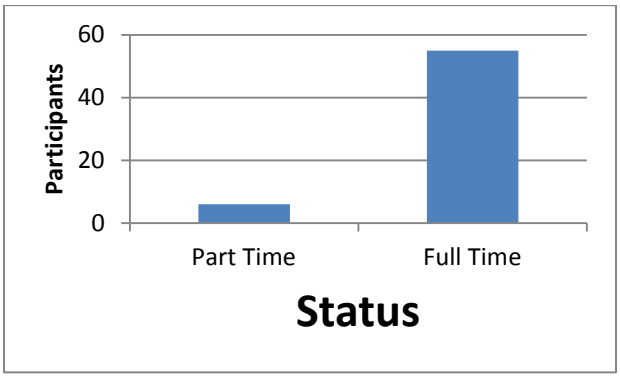


Figure 14 – Participant Part/Full Time Status Distribution

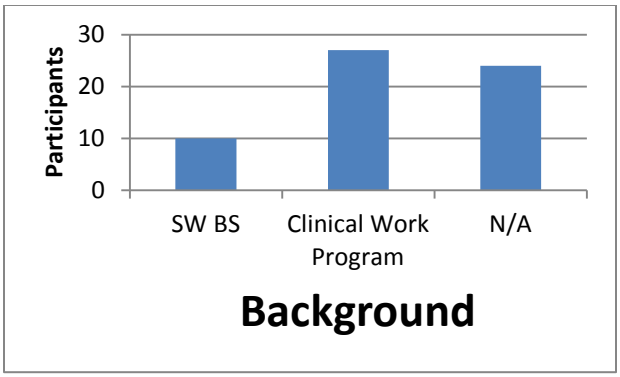


Figure 15 – Participant Educational Background Distribution

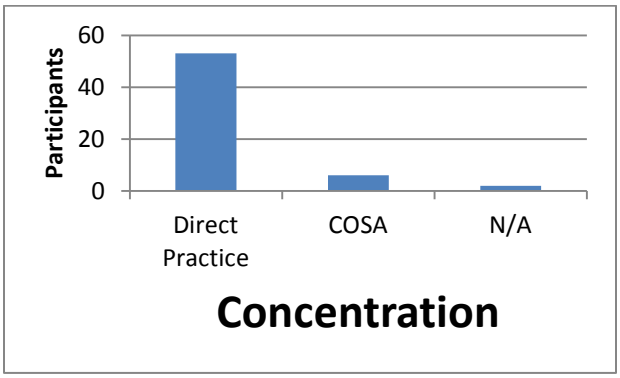


Figure 16 – Participant Field of Study Concentration Distribution

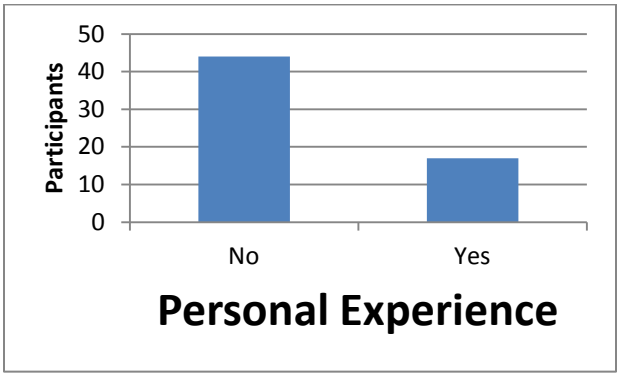


Figure 17 – Participant Personal Experience with Schizophrenia Distribution

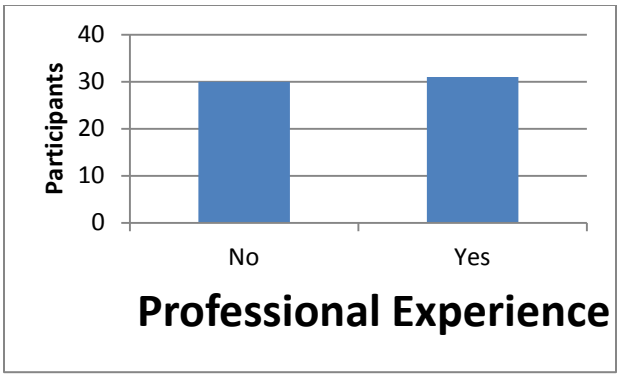


Figure 18 – Participant Professional Experience with Schizophrenia Distribution

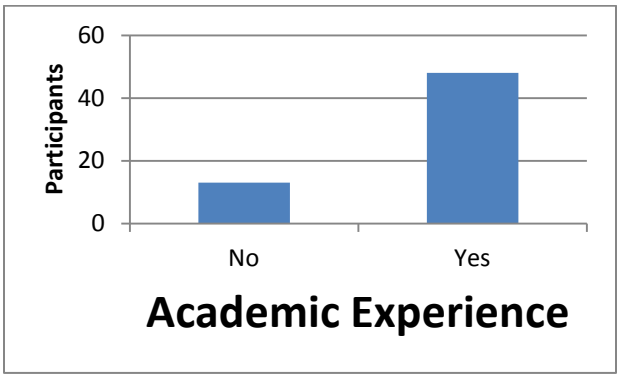


Figure 19 – Participant Academic Experience with Schizophrenia Distribution

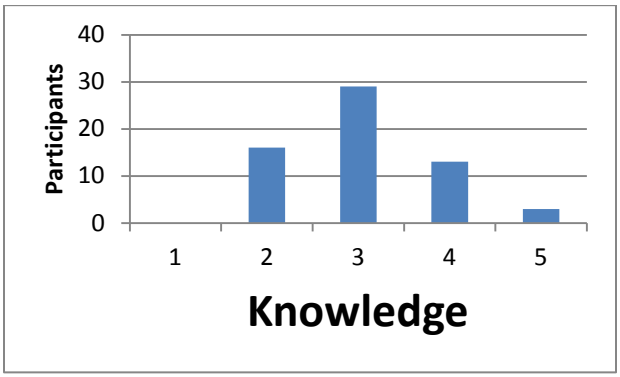


Figure 20 – Participant Knowledge of Schizophrenia Distribution

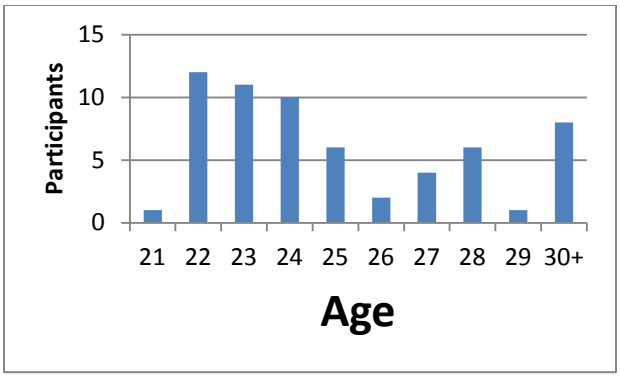


Figure 21 – Participant Age Distribution

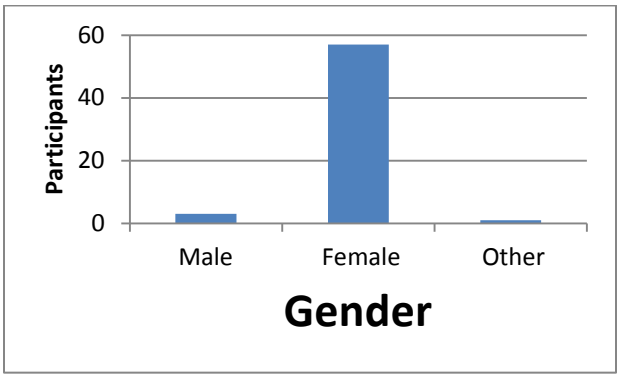


Figure 22 – Participant Gender Distribution

5.2.5 Hypothesis 1 Results

Hypothesis 1 referred to the potential difference between groups of the mean intervention judgement accuracy. Intervention judgement accuracy is calculated by adding the number of times a subject correctly identified that a message did or did not warrant an intervention response as compared to the gold standard, divided by the number of messages they were exposed to. Hypothesis 1 is restated below for clarity.

> $H_{1\text{-}0}$: There is no statistically significant difference in the mean intervention judgement accuracy with respect to the gold standard made by participants between the treatment and control cases.

> $H_{1\text{-}1}$: There is a statistically significant difference in the mean intervention judgement accuracy with respect to the gold standard made by participants between the treatment and control cases.

The gold standard here refers to the classification made in Part 1 of the study. If a message from the intervention corpus was judged by a participant to be in need of an intervention, this was considered a correct response. Likewise, if a message from the non-intervention corpus was judged to not need an intervention response, this was considered correct as well. Table 20 gives descriptive statistics for accuracy and Figure 23 shows the distribution of accuracy by participant and separated by case.

Table 20 – Descriptive Statistics for Accuracy by Case

Case	N	Minimum	Maximum	Mean	Std. Deviation
Control	30	.50	.78	.6260	.07722
Treatment	31	.44	.84	.6194	.09114

The performance of the two groups in terms of accuracy was very similar. The control and treatment group means differ by less than 1%. The mean of the control group accuracy being greater than the treatment group accuracy was a surprising result. An original expected result was the reverse, that the treatment group's mean accuracy would be greater. The treatment group did have a greater standard deviation as well as a lesser minimum and greater maximum for accuracy.

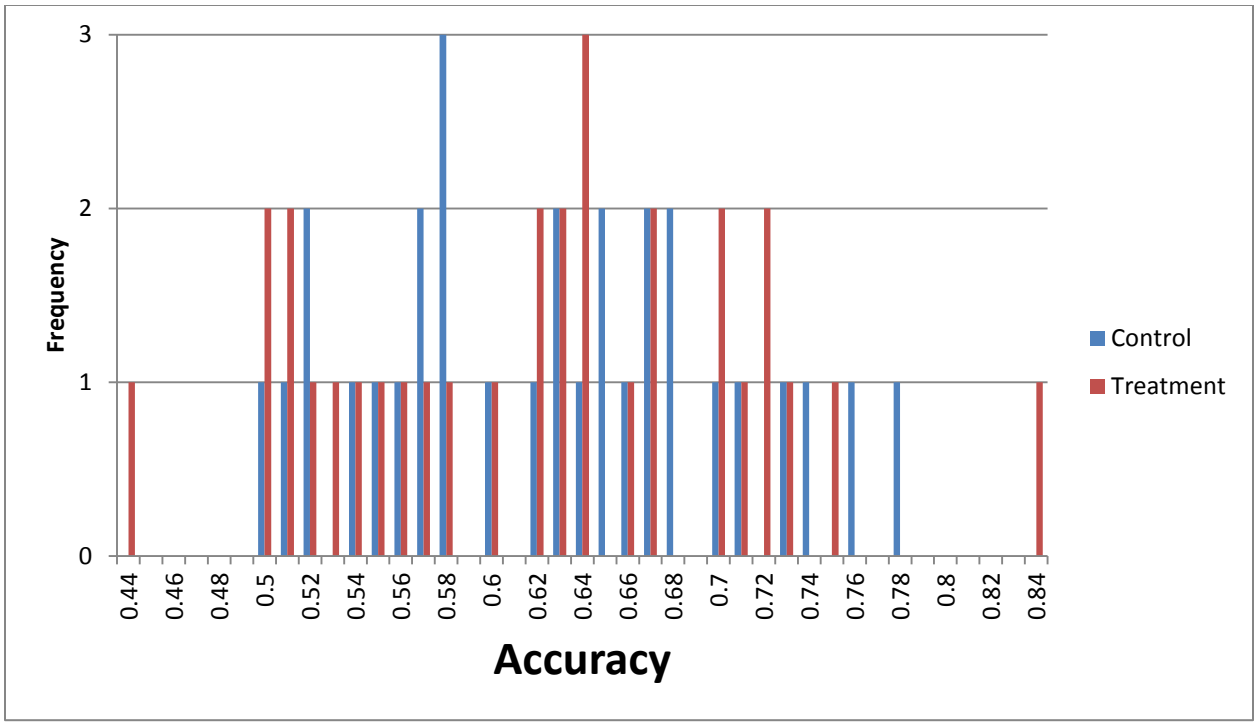


Figure 23 –Accuracy Distribution by Participant and Case

Figure 24 shows for each group the percent of accurate judgements for intervention messages, middle messages, and other non-intervention messages for each participant. The control case (odd numbers) participants are on the left half of the bar graph and the treatment case (even numbers) participants are on the right side.

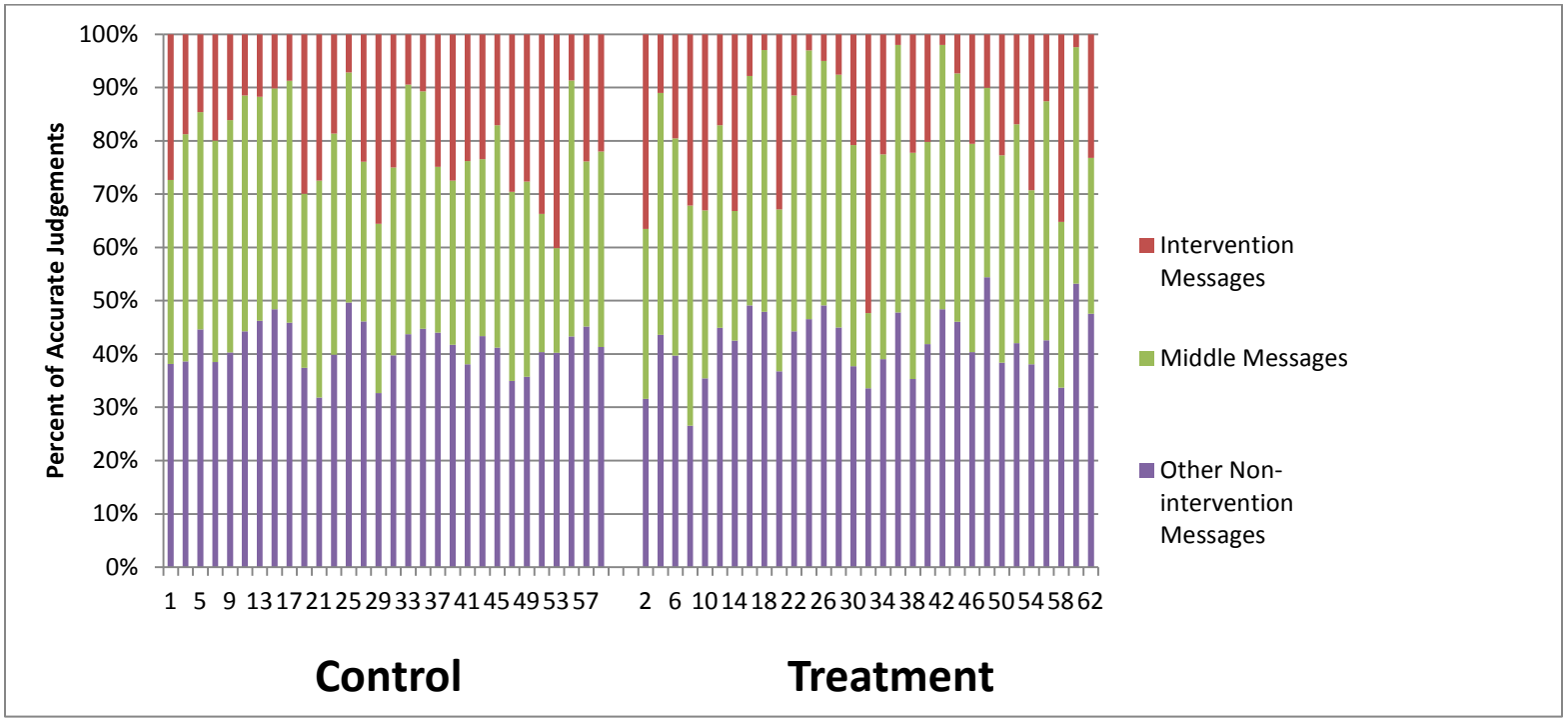


Figure 24 – Percentage of Accurate Judgements by Message Type and Participant

The results of the 1-way BS ANOVA test show that there is no significant difference between the two groups for accuracy, and therefore the null hypothesis has failed to be rejected, $F(1,59) = .094, p = .760, \eta_p^2 = .002$.

5.2.5.1 Excluding "Middle Messages"

This statistical analysis was also conducted on only the responses which were either intervention messages with rating of 8 or 10, or non-intervention messages with rating of 0. Table 21 shows the descriptive statistics for these messages with respect to accuracy.

Table 21 – Descriptive Statistics for Non-Middle Messages Accuracy by Case

Case	N	Minimum	Maximum	Mean	Std. Deviation
Control	30	.46	.77	.6054	.09179
Treatment	31	.40	.85	.5955	.10708

The results of the 1-way BS ANOVA test show that there is no significant difference between the two groups for accuracy, and therefore the null hypothesis has failed to be rejected, $F(1,59) = .149, p = .700, \eta_p^2 = .003$.

5.2.5.2 Truncated Messages

Tables 22 and 23 show the descriptive statistics for complete and truncated messages with respect to accuracy.

Table 22 – Descriptive Statistics for Complete Messages Accuracy by Case

Case	N	Minimum	Maximum	Mean	Std. Deviation
Control	30	.55	.84	.6783	.07921
Treatment	31	.46	.87	.6699	.09599

Table 23 – Descriptive Statistics for Truncated Messages Accuracy by Case

Case	N	Minimum	Maximum	Mean	Std. Deviation
Control	30	.31	.70	.5094	.10849
Treatment	31	.27	.77	.5043	.14603

For complete messages, there were no significant differences between the groups for accuracy, $F(1,59) = .140, p = .709, \eta_p^2 = .002$. For truncated messages, there was also no significant difference between the groups for accuracy, $F(1,59) = .023, p = .879, \eta_p^2 = .002$,

5.2.5.3 Discussion

The results of this hypothesis are disappointing. The visualization tool that was designed and built for this study was conceived initially as a means to improve moderator accuracy in a situation where context was limited. Providing the participants with more information was expected to have some sort of impact on the resultant performance. However, this is not borne out by the data. However, accuracy seemed to be higher for complete messages compared to truncated messages.

Given these results, it could be that any one of the underlying presuppositions of the experimental design were incorrect. For example, it might be that a trend line of sentiment score values is a very poor substitute for the full context of a message. Alternatively, it might be that no amount of visual information can make up for the richness of understanding obtained by reading the contextual messages surrounding any one message. Subjective feedback from the treatment case subjects on the usefulness of the visualizations is presented and discussed in section 5.2.9.

5.2.6 Hypothesis 2 Results

Hypothesis 2 referred to the potential difference between groups of the mean confidence ratings. Confidence ratings were reported by the participants for each message they were exposed to. These are separate from overall confidence which was reported by participants in the exit survey. Hypothesis 2 is restated below for clarity.

> $H_{2\text{-}0}$: There is no statistically significant difference in the mean confidence ratings made by participants for each intervention judgement between the treatment and control cases.

> $H_{2\text{-}1}$: There is a statistically significant difference in the mean confidence ratings made by participants for each intervention judgement between the treatment and control cases

Table 24 gives descriptive statistics for confidence and Figure 25 shows the distribution of confidence by participant and separated by case.

Table 24 – Descriptive Statistics for Confidence by Case

Case	N	Minimum	Maximum	Mean	Std. Deviation
Control	30	3.04	4.87	4.0040	.35175
Treatment	31	3.30	4.70	4.0019	.42313

Again the means of the two groups were very similar. The two groups have almost identical means. The treatment group had a greater standard deviation but the control group had a lesser minimum and greater maximum for confidence.

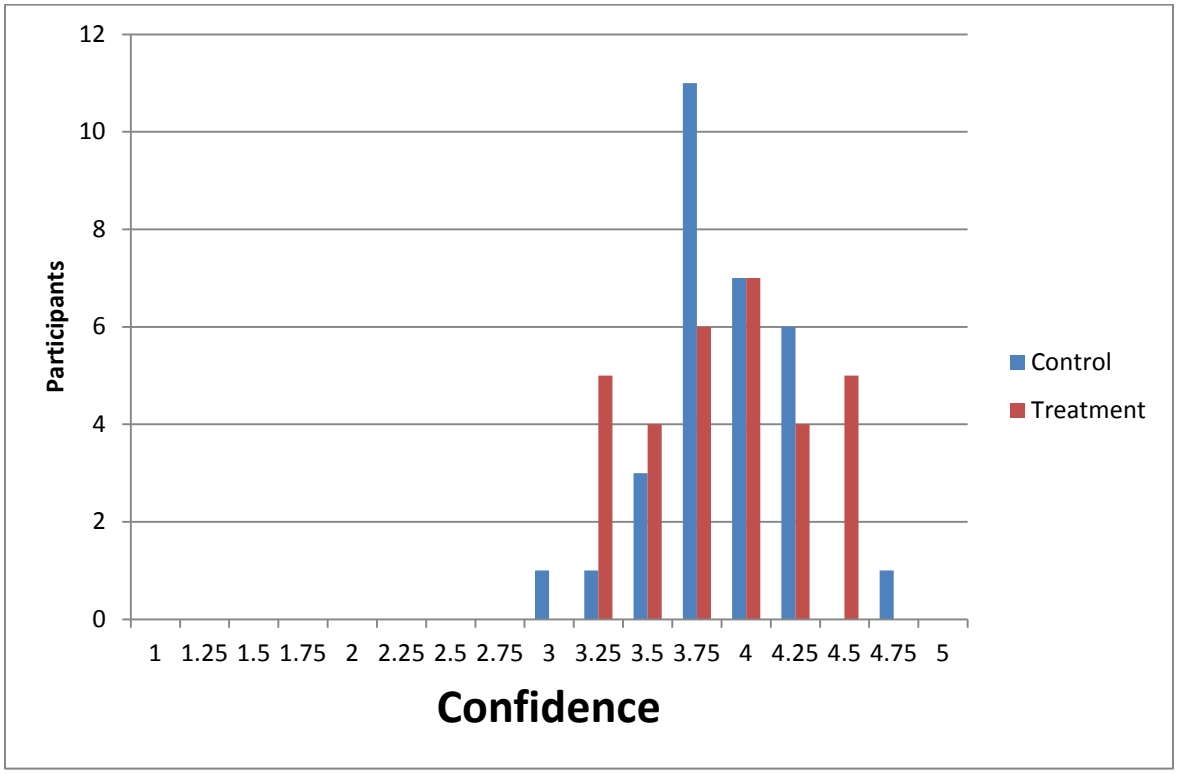


Figure 25 –Confidence Distribution by Participant and Case

The results of the 1-way BS ANOVA test show that there is no significant difference between the two groups for confidence, and therefore the null hypothesis has failed to be rejected, $F(1,59) = .000, p = .984, \eta_p^2 = .000$.

5.2.6.1 Excluding "Middle Messages"

This statistical analysis was also conducted on only the responses which were either intervention messages with rating of 8 or 10, or non-intervention messages with rating of 0. Table 25 shows the descriptive statistics for these messages with respect to confidence.

Table 25 – Descriptive Statistics for Non-Middle Messages Confidence by Case

Case	N	Minimum	Maximum	Mean	Std. Deviation
Control	30	3.05	4.87	3.9834	.36051
Treatment	31	3.27	4.69	3.9914	.41962

The results of the 1-way BS ANOVA test show that there is no significant difference between the two groups for confidence, and therefore the null hypothesis has failed to be rejected, $F(1,59) = .006, p = .937, \eta_p^2 = .000$.

5.2.6.2 Truncated Messages

Tables 26 and 27 show the descriptive statistics for complete and truncated messages with respect to confidence.

Table 26 – Descriptive Statistics for Complete Messages Confidence by Case

Case	N	Minimum	Maximum	Mean	Std. Deviation
Control	30	3.07	4.84	4.0873	.35675
Treatment	31	3.29	4.74	4.0739	.42185

Table 27 – Descriptive Statistics for Truncated Messages Confidence by Case

Case	N	Minimum	Maximum	Mean	Std. Deviation
Control	30	2.97	4.96	3.8353	.43367
Treatment	31	2.96	4.71	3.8451	.47693

For complete messages, there were no significant differences between the groups for confidence, $F(1,59) = .018, p = .894, \eta_p^2 = .000$. For truncated messages, there was also no significant difference between the groups for confidence, $F(1,59) = .007, p = .933, \eta_p^2 = .000$.

5.2.6.3 Discussion

Despite having more information at their disposal, the groups had no real difference in confidence in their judgements. The effect size is zero. It may be that the amount of information does not change a person's confidence in making a decision, but that seems to be intuitively

wrong and also the opposite of what is supported by the literature of various fields [109][110][111][112][113]. Additionally, confidence seemed to be higher for complete messages compared to truncated messages.

5.2.7 Hypothesis 3 Results

Hypothesis 3 referred to the potential difference between groups of the mean time taken per message. Time for each message was measured in milliseconds from when the message was displayed on screen until the "Save and Next" button was clicked. Units have been converted to seconds for readability. Hypothesis 3 is restated below for clarity.

> $H_{3\text{-}0}$: There is no statistically significant difference in the mean elapsed time for each intervention judgement between the treatment and control cases.

> $H_{3\text{-}1}$: There is a statistically significant difference in the mean elapsed time for each intervention judgement between the treatment and control cases.

Table 28 gives descriptive statistics for time and Figure 26 shows the distribution of time by participant and separated by case.

Table 28 – Descriptive Statistics for Time (s) by Case

Case	N	Minimum	Maximum	Mean	Std. Deviation
Control	30	8.16	26.34	12.7554	3.79366
Treatment	31	7.57	36.38	14.7553	5.41620

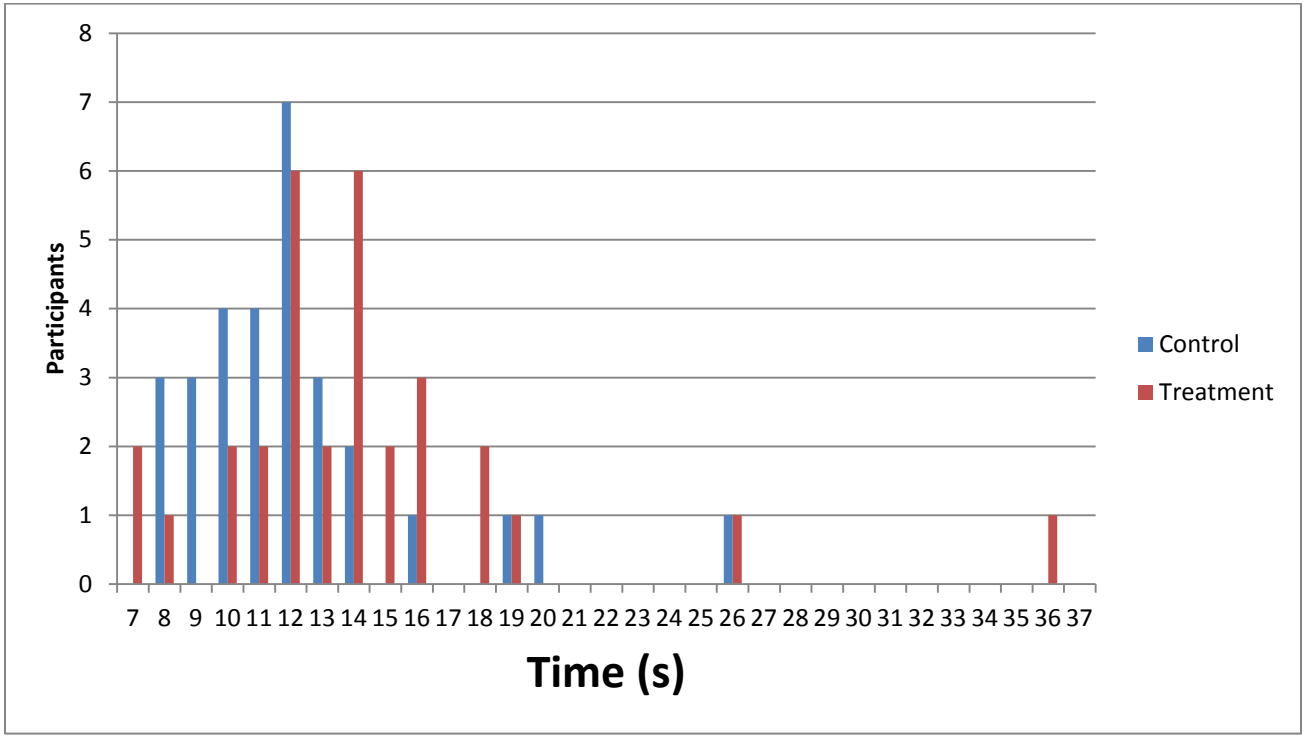


Figure 26 –Time (s) Distribution by Participant and Case

The results of the 1-way BS ANOVA test show that there is no significant difference between the two groups for time, and therefore the null hypothesis has failed to be rejected, $F(1,59) = 2.773, p = .101, \eta_p^2 = .045$.

5.2.7.1 Excluding "Middle Messages"

This statistical analysis was also conducted on only the responses which were either intervention messages with rating of 8 or 10, or non-intervention messages with rating of 0. Table 29 shows the descriptive statistics for these messages with respect to time.

Table 29 – Descriptive Statistics for Non-Middle Messages Time by Case

Case	N	Minimum	Maximum	Mean	Std. Deviation
Control	30	8.32	25.97	12.8457	3.78662
Treatment	31	7.58	36.75	14.8778	5.52254

The results of the 1-way BS ANOVA test show that there is no significant difference between the two groups for time, and therefore the null hypothesis has failed to be rejected, $F(1,59) = 2.791, p = .100, \eta_p^2 = .045$.

5.2.7.2 Truncated Messages

Tables 30 and 31 show the descriptive statistics for complete and truncated messages with respect to time.

Table 30 – Descriptive Statistics for Complete Messages Time by Case

Case	N	Minimum	Maximum	Mean	Std. Deviation
Control	30	7.07	24.90	11.1067	3.72294
Treatment	31	6.54	32.86	12.8688	5.00585

Table 31 – Descriptive Statistics for Truncated Messages Time by Case

Case	N	Minimum	Maximum	Mean	Std. Deviation
Control	30	10.01	29.55	16.3012	4.11890
Treatment	31	9.42	48.19	19.0607	7.01670

For complete messages, there were no significant differences between the groups for time, $F(1,59) = 2.421, p = .125, \eta_p^2 = .039$. For truncated messages, there was also no significant difference between the groups for time, $F(1,59) = 3.479, p = .067, \eta_p^2 = .056$.

5.2.7.3 Discussion

These results are somewhat in line with what was expected. Time was expected to increase overall for the treatment group due to the increased amount of information to be absorbed while viewing each message. Despite this, there is still no statistically significant difference between the groups, and the effect size is small.

5.2.8 Exit Survey Feedback

The exit survey questioned the participants for their feedback on a variety of topics relating to their experience during their sessions. The exit survey for the control group was shorter than the treatment group's because the latter included questions relating to the ADAS tool which the control group was not exposed to. The exit survey contained free-response questions as well as multiple-choice questions.

There were 6 multiple choice feedback questions. Two of these were presented to both cases and four were presented to the treatment case only. Figures 27-32 present the result distributions.

In Figure 27, the distribution for overall confidence in judgements is shown. This question was aimed at getting a single overarching degree of confidence for each participant with respect to the overall task. Table 32 shows descriptive statistics for the responses.

Table 32 – Descriptive Statistics for Overall Judgement Confidence by Case

Case	Mean	Std. Deviation	N
Control	3.3333	.60648	30
Treatment	3.4194	.67202	31

The results of a 1-way BS ANOVA test for this measure resulted in there being no significant difference found between the responses of the two groups, $F(1,59) = .275, p = .602, \eta_p^2 = .005$.

Figure 32 shows the distribution of the other question asked to both groups, whether longer or shorter messages had more impact on judgement making confidence. The definition of 'longer' and 'shorter' were deliberately left vague and for the participants to determine for themselves. The majority of both cases indicated they thought longer messages had more impact on their confidence. The results of a 1-way BS ANOVA test resulted in there being no significant difference found between the responses of the two groups, $F(1,59) = .384, p = .538, \eta_p^2 = .006$.

Figure 28 shows the distribution for the responses to the question of overall confidence in the visualizations. This was meant to represent the confidence the participants had in the efficacy of the visualization and in the visualization's ability to accurately represent reality. This question was only asked of the treatment group. Most participants indicated moderate confidence in the visualizations (mean = 3.13).

Figures 29 and 30 illustrate that the majority of the treatment group reported that the sentiment trend plot had more impact on their confidence and helped their decision making more than the trigger word highlighting.

The question relating to Figure 31 asked if judgement confidence was impacted by messages which had fewer than 10 previous messages to display in the visualization. Interestingly, the majority responded no. This result seems to agree with the results of the accuracy and confidence tests in hypotheses 1 and 2. Specifically, if there was no difference between the groups for accuracy and confidence, then it would imply dearth of message history would have no impact either.

Figure 27 – Overall Confidence in Judgements Distribution by Participant and Case

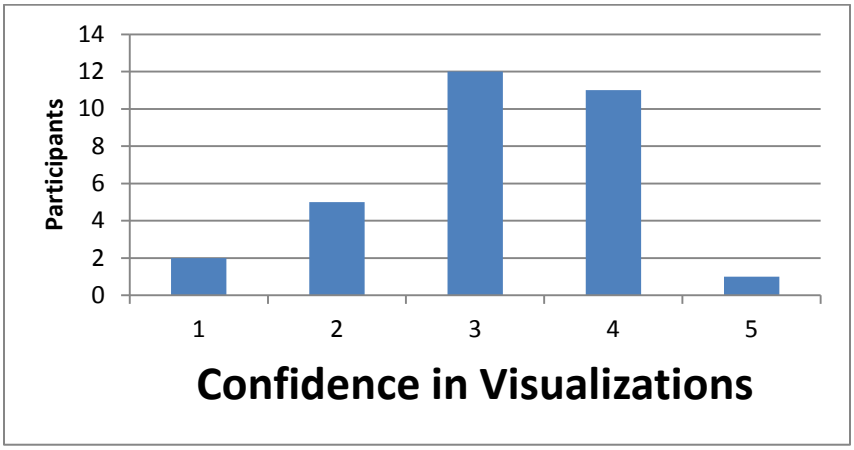


Figure 28 – Overall Confidence in Visualizations Distribution by Participant (Treatment Case Only)

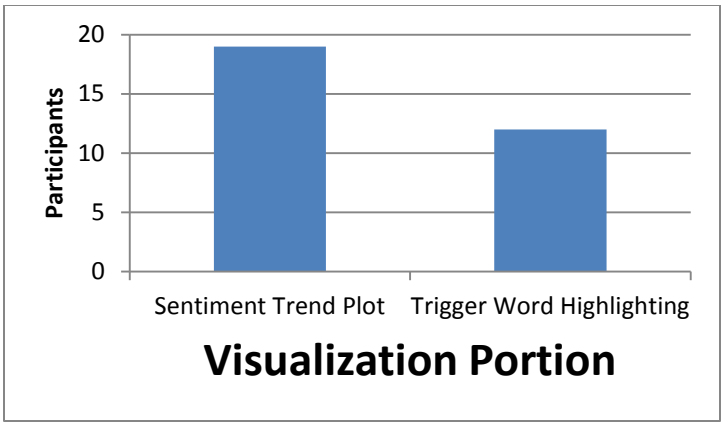


Figure 29 – Distribution of Which Part of Visualization Impacted Confidence in Judgements Most (Treatment Case Only)

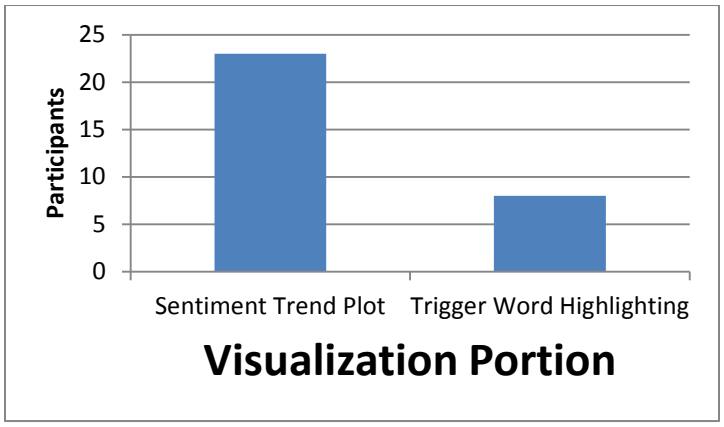


Figure 30 – Distribution of Which Part of Visualization Helped Making Judgements Most (Treatment Case Only)

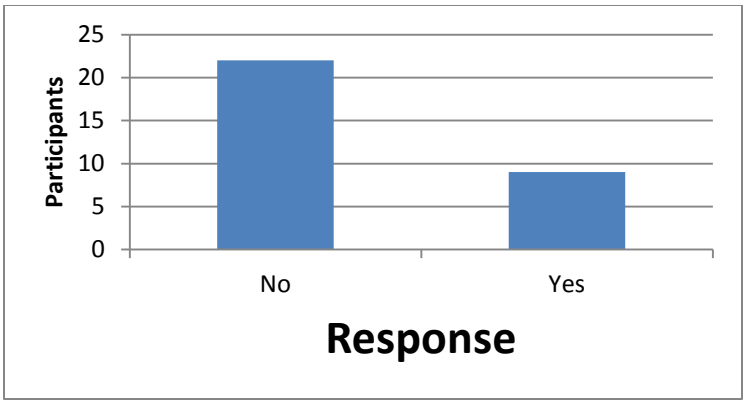


Figure 31 – Distribution of If Judgement Was Impacted by Messages with Fewer than Ten Prior Messages (Treatment Case Only)

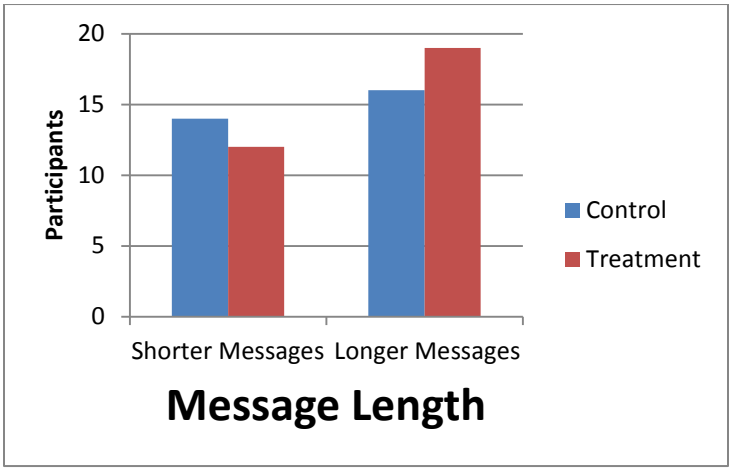


Figure 32 – Distribution of Which Length of Message Impacted Confidence in Judgements Most by Case

5.2.9 Exit Survey Free Response Feedback

There were three free response questions on the exit survey. Two of them were posed only to the treatment group because they were related to the ADAS visualizations. The following sections are organized by question.

5.2.9.1 Extra Information

The first free response question was posed to both the control and treatment cases. The question as worded in the survey was "What kind of information would have made this task easier for you?" This question accomplished different goals between the cases. For the treatment group, this was designed to get feedback relating to what should have been included in the ADAS visualizations but was not, and for the control group it was designed to hopefully elicit the kind of information presented to the treatment group. Out of 61 participants, two from the control group and three from the treatment group provided no feedback for this question.

From the control group, responses varied but a main theme was a desire for more context to be presented for each message. In fact, this was a prevailing theme in the treatment group responses as well. A typical control group response:

> "Knowing if these messages where *[sic]* comments to another message, and what that message was. Maybe even seeing replies to the messages."

A typical treatment group reply expressed the same desire explicitly:

> "I would like to see the messages in context. This would include (1) the previous messages in the thread, (2) previous messages in other threads, and (3) following messages in all threads…"

Context in the form of prior messages being displayed in their entirety was not provided as part of the research design in order to test the impact of showing the sentiment score. Messages following the one being shown were not provided for the same reason and also because it would not represent the real-life situation faced by moderators, who read a message in real time and need to decide if an intervention is to be made or not in that moment.

5.2.9.2 Visualization Likes

The second free response question was only seen by the treatment group. It was worded as "What did you like the most about the visualizations?" Despite the desire for additional context discussed above, the general trend of feedback for this question was positive and supportive of the visualization's usefulness. Three participants did not answer this question. Some typical responses include:

> "The visualizations assisted with knowing the person's pattern…being positive or negative"

> "The line plot allowed me to assess how often the person was writing...if I noticed…multiple days between messages, then more recently…I imagined that they may be experiencing some sort of episode…"
>
> "I liked the visualizations because it helped provide context for how that person had been feeling…"

Other responses contrasted the usefulness of the sentiment plot versus the trigger word highlighting. In most cases, the sentiment plot was reported as being most useful. However, despite this positive feedback claiming the impact it had on their task, there was no impact on the performance of this group large enough to be considered significant.

5.2.9.3 Visualization Dislikes

The last free response question was also only seen by the treatment group. It was worded as "What did you not like the most about the visualizations?" For this question the feedback was mainly concerned with the uselessness of the trigger word highlighting. Participants felt that it was confusing, missing the point, not relevant, etc. Again, three participants did not answer this question. Typical responses include:

> "…Trigger words were not very helpful. They did not harm my judgment or make it harder but I hardly ever used them."

"The highlighted words didn't help me much in contemplating an intervention."

"Some red words were distracting"

The dislike for the trigger words here is also expressed in the multiple choice questions, where most participants found the plot had the most helpfulness and impact on confidence. Responses also talked about difficulty interpreting the sentiment plot timeline or not being able to actually read past messages.

5.2.10 Non-Intervention Messages with Rating >0 "Middle Messages"

As described in section 5.1.4.3, there was a subset of messages presented during the study which had been given a rating of greater than zero and combined with other non-intervention messages with a rating of zero. This section explores the performance of the participants on these "middle messages".

Throughout the run of part 2 of the study, participants were shown 6100 messages, 100 per session. Of the 100 messages in each session, 50 were to come from the non-intervention group and 50 were to come from the intervention group. Therefore, of the 6100 total messages shown, 3050 came from the intervention group, and 3050 came from the non-intervention group.

In the 3050 messages shown from the non-intervention group, 2374 (77.8%) had a rating of 0 and 676 (22.2%) had a rating of >0. Figure 33 shows a stacked distribution of the ratings of all the messages shown to each participant.

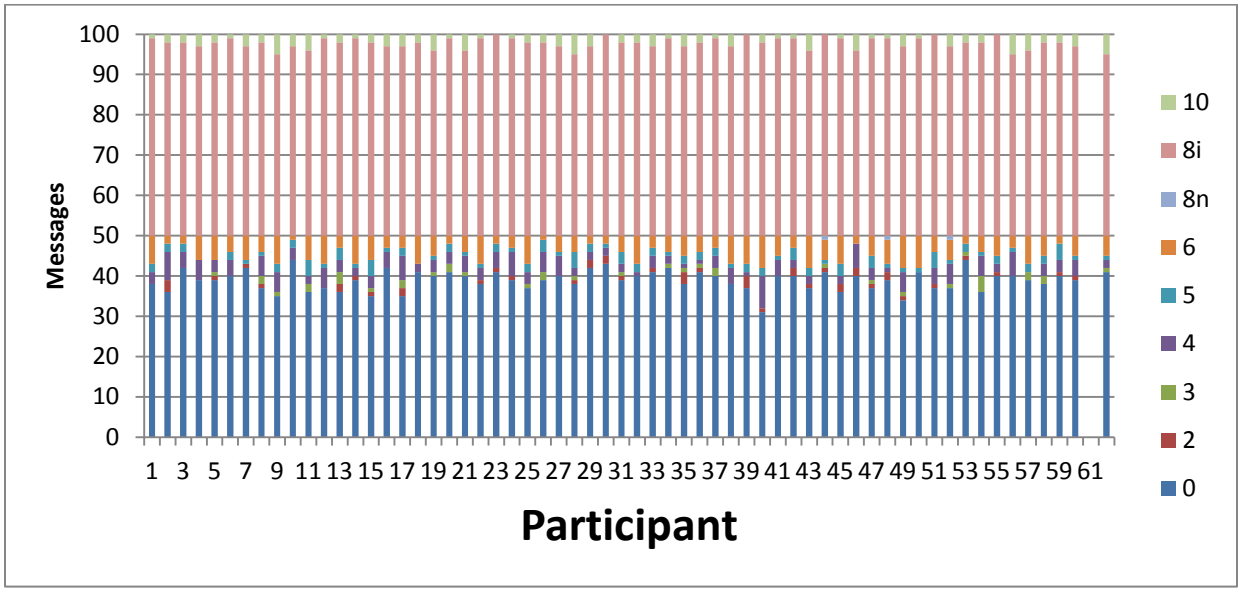


Figure 33 – Distribution of Ratings Shown to Each Participant

Number 61 is empty because the 61st participant was assigned to the treatment group which received even numbers, and was given number 62. The legend on the right of the figure color codes the various ratings. The ratings 8i and 8n correspond to those messages which were rated as 4/4 vs 5/3, the latter of which were part of the non-intervention message group. In general, the figure illustrates that most intervention messages have a rating of 8 (pink) and most non-intervention messages have a rating of 0 (blue). Table 33 shows the raw counts and relative percentage for each rating. Figure 34 "zooms into" the previous figure and shows the distribution of only the "middle messages", i.e. non-intervention messages with rating greater than 0.

Table 33 – Raw Counts and Percentage for each Rating

Score	2	3	4	5	6	8n	Total
Count	47	40	193	107	286	3	**676**
Percent	6.95	5.92	28.55	15.83	42.31	0.44	**100**

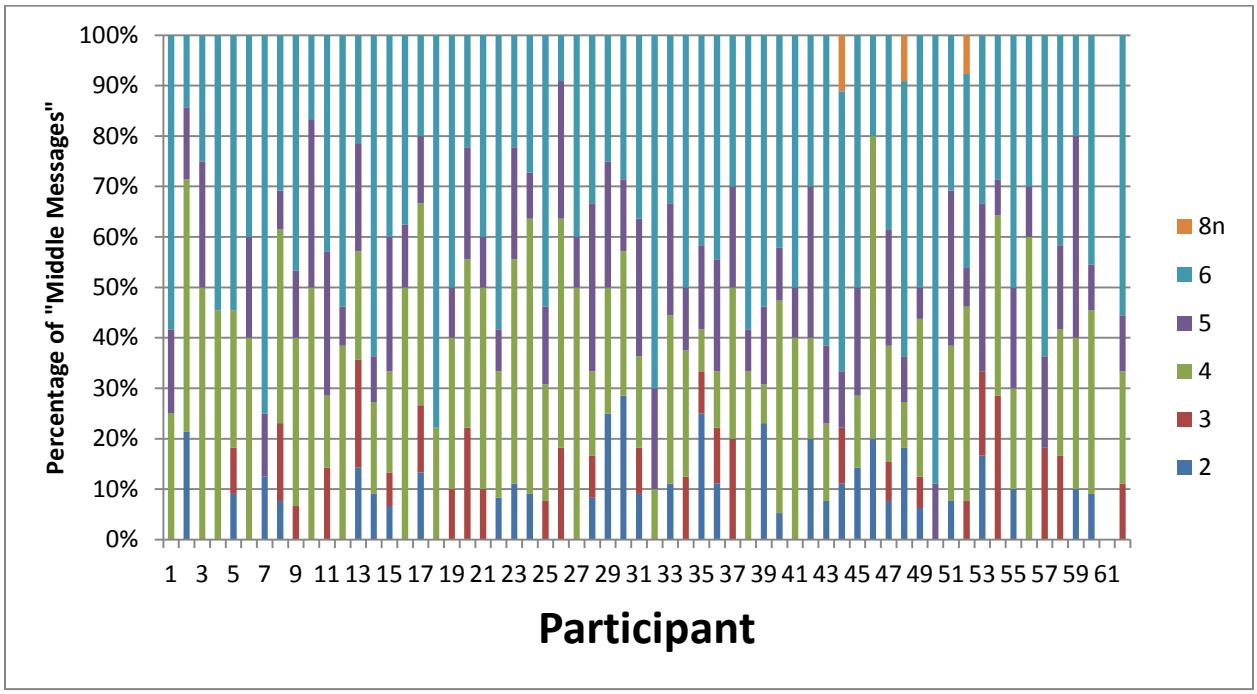


Figure 34 – Distribution of Ratings for “Middle Messages” by Participant

Most of the “middle messages” shown to participants had ratings of 4, 5, and 6. These accounted for 86.92% of all such messages shown. Figure 35 shows the distribution of judgement responses across “middle messages”. Most often, participants answered that these messages did not warrant an intervention response. The average rating of messages which received a “yes” response (intervention warranted) was 5.26 and the same for “no” responses (intervention not warranted) was 4.69.

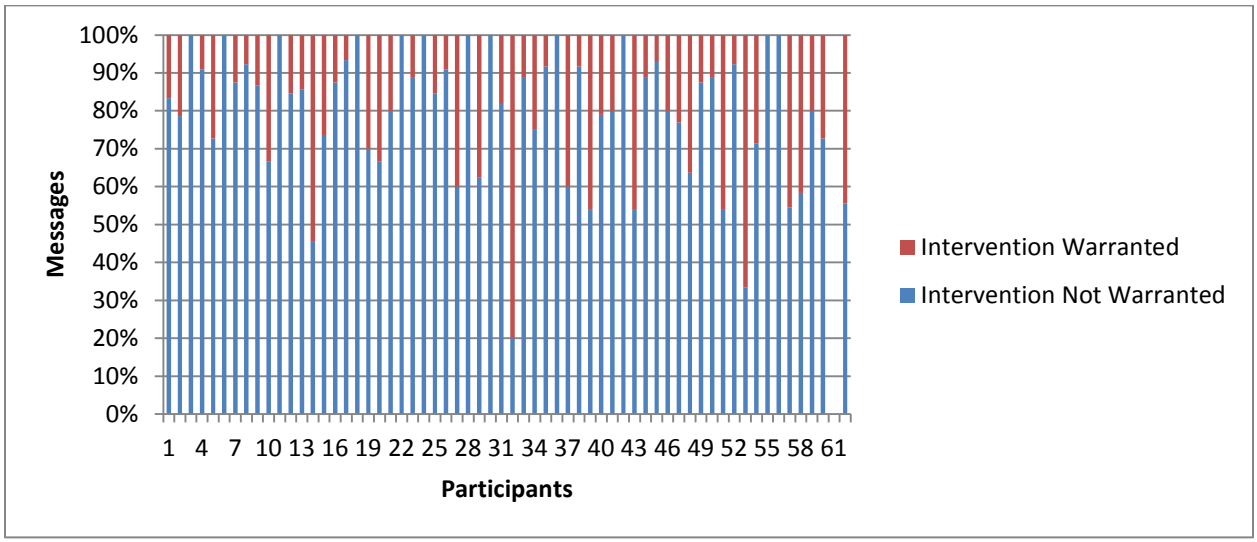


Figure 35 – Distribution of Judgement Responses for "Middle Messages" by Participant

5.2.11 Participant Field of Study Comparison

Due to the differing fields of study of the participants, this section explored the performance of the SSW participants against the participants from SHRS and SN. 52 of the participants came from SSW, with 8 from SHRS and 1 from SN. Table 34 presents data from the entry questionnaire separated by participants from SSW and otherwise.

Table 34 – Entry Questionnaire Comparison by Participant Field of Study

Field	Control N	Treatment N	Average Age	Average Experience in Years	Average Knowledge of Schizophrenia
SSW	27	25	25.3	2.42	2.98
SRHS+SN	3	6	23.22	1.22	3.44

The participants from SSW were about equally distributed between the control and treatment groups. Non SSW participants were skewed towards the treatment group, but this is only due to chance during the group assignment, which was made without regards to any participant attributes other than when they scheduled their sessions.

Non SSW participants were younger and had fewer years of experience in their fields of study on average. However, the non SSW participants reported they had more knowledge of schizophrenia on average than SSW participants.

Table 35 shows a breakdown of average performance by SSW and non SSW fields within the control and treatment groups for accuracy, confidence, and time.

Table 35 – Performance of Participant Fields of Study within Control and Treatment Groups

	Control		Treatment	
	SSW	Non SSW	SSW	Non SSW
Average Accuracy	0.62	0.69	0.62	0.6
Average Confidence	4.03	3.78	4.02	3.93
Average Time(s)	12.78	12.51	14.85	14.34

In general, the average performance of SSW and non SSW participants were similar to each other. In the control group, SSW participants were slightly less accurate, had slightly higher confidence, and took slightly more time than non SSW participants on average. In the treatment group, the same was true for confidence and time on average, but SSW participants had slightly higher accuracy on average than non SSW participants. There were no significant difference found in the performance between SSW and non SSW participants.

5.2.12 Message History

In section 3.6, a message relative history heuristic was introduced. The goal of this heuristic was to provide some way to rank each message's degree of relative history at the time it was posted to the DSW discussion forum. When a moderator sees a message for the first time, it might be that the author is either prolific or not prolific and the discussion forum might be active as a whole or inactive as a whole. Taking these two variables into account, the history heuristic is designed to produce a score based on the activity of a message's author compared to the activity of all the forum's authors. Table 36 shows descriptive statistics for the history heuristic scores for all messages and also broken down by message type. Figure 36 shows a distribution of the history heuristic score by message type. Outliers are excluded for readability.

Table 36 – Descriptive Statistics for History Heuristic Score

	N	Min	Max	Average	Median	Std. Deviation
Non Intervention Messages	4682	0	0.2222	0.0218	0.0108	0.0255
Intervention Messages	306	0	0.0960	0.0125	0.0058	0.0166
All	4988	0	0.2222	0.0212	0.0105	0.0252

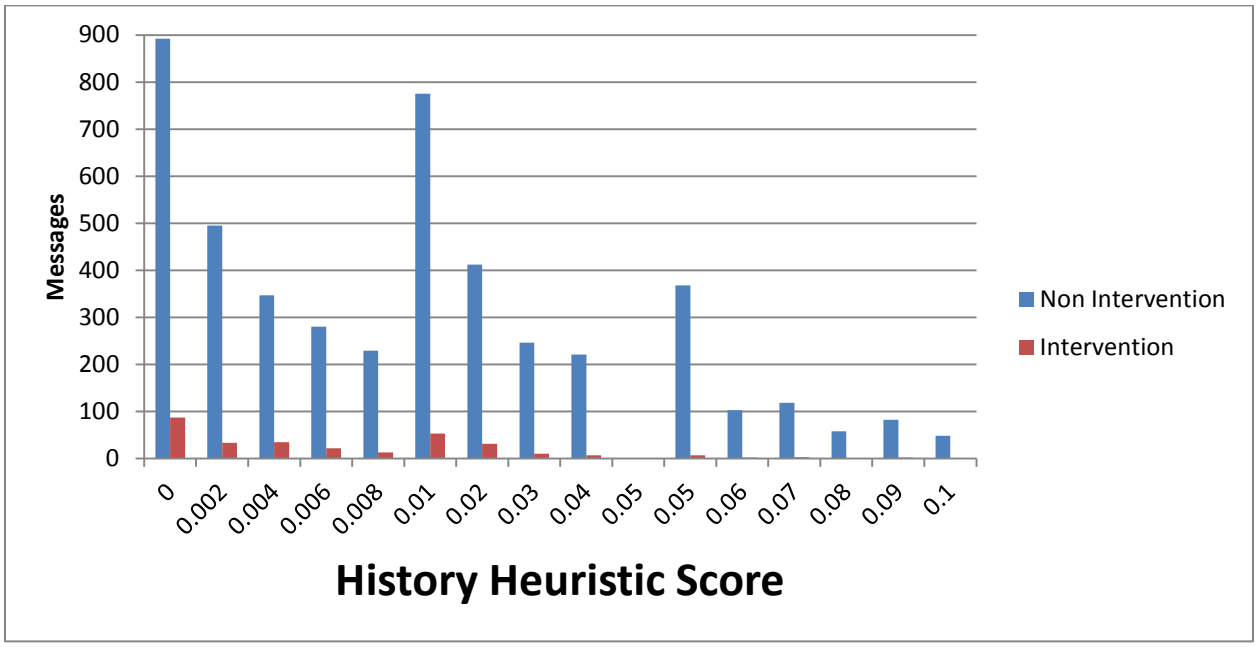


Figure 36 – Distribution of History Heuristic Score by Message Type (Outliers Excluded)

Due to outliers, the average for history score is unrepresentative of the distribution of scores. Therefore, the distribution has been split into low and high history segments based on the median score. The performance of the control and treatment groups on high and low history messages in terms of accuracy, confidence, and time is examined in the following section.

5.2.12.1 Impact on Performance

Table 37 shows the average accuracy, average confidence, and average time for intervention and non-intervention messages shown to the treatment and control groups divided into the low and high history segments.

Table 37 – History Segment Comparison for Accuracy

		Accuracy		Confidence		Time(s)	
		Low History	**High History**	**Low History**	**High History**	**Low History**	**High History**
Control	**Intervention**	0.451	0.404	3.769	3.82	13.865	14.368
	Non Intervention	0.816	0.833	4.218	4.208	11.581	11.193
Treatment	**Intervention**	0.376	0.391	3.786	3.868	16.233	16.482
	Non Intervention	0.849	0.859	4.165	4.189	13.375	12.95

There are slight differences between low and high history message segments for each performance measure, but there were no significant differences found. The main impact of message history seems to have been subjective. As described in section 5.2.9, the participants most often desired additional context around the messages they were being shown beyond the sentiment trend plot. Even though some messages had a higher history score than others, none of the surrounding context that would be available for the higher score messages was provided to the participants in any case. This might be one reason why history score had no significant impact on performance.

5.2.13 Participant vs Classifier Performance Comparison

This section subjectively compares the performance of the study participants in Part 2 to the performance of the classifiers in Part 1. The tasks these groups performed were distinct but the comparison being sought is the relative proclivity to make interventions by the Part 2 participants versus how those messages were classified in Part 1.

Table 38 shows the number of times messages of each rating made by the classifiers were shown in Part 2 and also the percentage of the time that those messages were deemed to warrant an intervention, split by case. Participants in both cases were most apt to judge that interventions were warranted for messages with higher ratings. In both groups, participants were in general less inclined to decide an intervention was warranted for messages with lower ratings.

Table 38 – Counts and Percentage of Judgements Warranting Interventions of Messages Shown in Part 2 by Rating and Case

	Count		Intervention Deemed Warranted	
Rating	Control	Treatment	Control	Treatment
2	27	20	7.41%	5.00%
3	22	18	18.18%	5.56%
4	85	108	14.12%	8.33%
5	64	43	21.88%	25.58%
6	145	141	29.66%	26.24%
8n	0	3	-	33.33%
8i	1433	1485	41.17%	36.90%
10	67	65	76.12%	72.31%

Furthermore, the participants in both groups were less inclined than the classifiers to decide that an intervention was necessary. For example, the classifiers were decided that all the messages they rated 10 were most in need of an intervention. In 132 instances (67 control and 65 treatment) of 10 rated messages being shown, participants decided that an intervention was needed only 76% of the time in the control group and 72% of the time in the treatment group. A possible explanation is that the classifiers having worked as the original DSW moderators had a better sense of the history of the messages they saw when making their ratings. The lack of this extra information could partly explain the more cautious decision making of the participants.

5.2.14 Learning Effect

This section examines whether there was any significant difference for the treatment group's performance with respect to accuracy, confidence, and time when considering only the first and last twenty messages shown during each session. This is to determine whether there was a

learning curve which each treatment group member endured when using the ADAS tool for the first time versus when they had been using it for some time already.

Statistical analyses were conducted for the three measures of accuracy, confidence, and time. The cases for these analyses were "beginning" and "end", referring to the average measures for each participant during the first and last 20 messages in a session. Tables 39, 40, and 41 show the descriptive statistics for these cases with respect to accuracy, confidence, and time respectively.

Table 39 – Descriptive Statistics for Beginning and End Accuracy

Case	N	Minimum	Maximum	Mean	Std. Deviation
Beginning	31	.45	.85	.6484	.11216
End	31	.05	.90	.5468	.23092

Table 40 – Descriptive Statistics for Beginning and End Confidence

Case	N	Minimum	Maximum	Mean	Std. Deviation
Beginning	31	3.30	4.95	4.0016	.41099
End	31	2.95	4.90	3.9387	.50128

Table 41 – Descriptive Statistics for Beginning and End Time

Case	N	Minimum	Maximum	Mean	Std. Deviation
Beginning	31	8.89	38.29	19.3722	7.00168
End	31	6.40	34.10	12.9977	5.36662

The results of 1-way BS ANOVA tests show that there is a significant difference between the two cases for accuracy, $F(1,61) = 4.857, p = .031, \eta_p^2 = .075$, and for time, $F(1,61) = 16.186, p = .000, \eta_p^2 = .212$. There was no significant difference for confidence, $F(1,61) = .292, p = .591, \eta_p^2 = .005$.

The ending case for the measure of accuracy had a lower mean than the beginning case. Therefore, the significant difference seen in the test results indicate that there was a reverse learning effect i.e. the treatment case participants were more accurate on average at the beginning of their sessions than at the end.

The ending case for the measure of time also had a lower mean than the beginning case, but this is in line with how a normal learning curve would operate i.e. there is more time taken at the beginning while the tool is being learned than at the end when there is more understanding.

CHAPTER 6: CONCLUSION

In this chapter, an outline for the vision of what the work in this dissertation could lead to is laid out in section 6.1 along with a discussion of the impact of the work completed in this dissertation in section 6.2. Last, directions of future work are explored in section 6.3.

6.1 VISION

This dissertation was conceived as a first step working toward a grand vision wherein individuals with cognitive disability can widely utilize online discussion forums to their full positive benefit. So far, such discussion forums have been part of only small-sized research studies accommodating perhaps a few hundred individuals and moderated keenly by full time experts. These moderators have relied on reading and consuming all the messages posted to the discussion forums as well as their expert experience and skills in order to have the context necessary to know when and where to judiciously make therapeutic interventions. When necessary, these interventions preserve the positive and organized nature of the discussion forum, but when unnecessary they break down the organic nature of the discussion on the forum.

If this sort of resource were to be available to the wider population of the cognitively disabled, these moderating requirements might well become prohibitively costly in terms of both

manpower and money. There might not be enough highly trained and skilled social work or counseling professionals employable as moderators to scale with the size of the number of people who would want to utilize these sorts of resources and also it might be difficult to compensate this number of professionals commensurate with their education and experience.

Therefore, it is envisioned that a sophisticated automated decision aiding system (ADAS) software tool could aid a smaller number of such professional moderators in identifying the sorts of messages which require their direct attention. Also, such a tool would be able to provide the necessary context for those messages in an easily digestible format with high fidelity without having to laboriously read all the messages surrounding them. With the aid of such an ADAS, the number of moderators might not need to scale linearly with the number of individuals utilizing an online discussion forum.

6.2 CONTRIBUTION AND IMPLICATIONS

The American physicist Richard Feynman is quoted as saying “If it doesn't agree with experiment, it's wrong.” It is also often remarked that a rigorous experiment with no significant results is just as valid and useful as one that does, and may well be more useful because by virtue of the ‘failure’, whatever was being tested can be ruled out of the pool of potential successful solutions.

To this end, the main contribution of dissertation is that it has successfully ruled out an ADAS configuration that has been shown to not have any significant impact on the accuracy, confidence level, or speed of moderating an online discussion board with low context.

The research question posed at the beginning of this study was:

> Can the automated analysis and visualization of an author's messaging behavior on a controlled access online social network discussion forum allow experts to moderate such forums more efficiently?

This question can now begin to be answered with the results of this study. In this case, the data suggests that the answer is no. Of course, this is only for the ADAS examined in this study. There is no way to know for certain ahead of time if this problem is intractable, but in the more than likely case that it is not, many directions for future research can be started from where this study left off.

6.3 FUTURE WORK

The directions for future work laid out in this section are understood to be described in ideal scenarios. With that in mind, future work relating to this study fall into one of a few broad categories, discussed in the following sections.

6.3.1 Participant Experience

The participants in this study were selected as a practical approximation of the professionals who are envisioned to be the target users of a successful ADAS for this task. It might be that the

difference in experience, understanding, or skill level between those who participated in this study and the professionals in the field of social work is a confounding factor. That is, the treatment and control cases would have done equally well at this task in any situation based on their lack of ability to perform the task at all.

If this is the case, it could be tested in an ideal experiment by acquiring an equally large sample of experienced professionals and reattempting this study. The time and recruiting costs for this would be considerable. The results would however give insight into the difference that experience and skill makes in the performance of this task.

6.3.2 ADAS Tools

The ADAS tool which was conceived and designed for this study was based on a few underlying assumptions such as the general usefulness of sentiment polarity in messages as a substitute for context. It could be that simply indicating that a message someone had written in the past was 'happy' or 'sad' distills out too much of the richness contained in natural language for a person to have any chance at performing better at making judgements on single messages than having no extra information at all. In this case a "back to the drawing board" approach might be best. Testing other forms of context-supplying ADAS tools such as automatic summarization could yield better results. In future work based off of this ADAS redesign, less experienced participants could again be used, in order to determine the point at which the context supply is high enough to produce a significant difference between groups.

6.3.3 Context Supply

It might be that the amount of context supplied by the ADAS in this study was too little to make any difference in the performance of the groups. The abstraction of context provided in the treatment interface caused the participants to not have access to many aspects of messages which were available to the DSW moderators such as names of message authors, posting history, and surrounding messages in the same topic thread. The distillation and abstraction of these contextual elements by the ADAS might always be inferior to the human ability to process natural language. Furthermore, some aspects of the treatment interface may have introduced noise rather than information, such as the trigger word highlighting, which might be why accuracy decreased over time from the beginning of each session to the end within the treatment group. This can be mitigated in the future by a more robust and iterative prototyping process. This study was meant to only test the ADAS tool and not to test the large body of literature which forms the theory of the effects of limiting information supply.

A larger experiment might also include multiple cases, from a control group similar to the one used for this study though to allowing all the context of each and every message to be shown and explored at length by the participants, just like the original DSW moderators, and separated by degrees. Performing the same task as in this study but with this greater number of groups would give some insight into the point at which there is some critical minimum of context being supplied that gives a group what they need to perform better. If the results of this sort of experiment show there is still no separation, then it might imply that this is an intractable problem in so far as human ability to process natural language is not able to be surpassed with the current state of technology.

6.3.4 Minimizing Error

This study was conceived, designed, and executed with a primary goal of maximizing the fidelity of the data collected and minimizing the possible sources of error which would undermine the same. Sometimes it can seem that a given situation no amount of preparation or caution can prevent every undesirable eventuality. Sanity checks of the data and system performance were carried out in good faith and all errors which were detected before data gathering took place were corrected speedily.

Nevertheless, this study was subject to preventable errors such as the truncation of a subset of the messages presented to participants. It may be that some other unknown or unobserved human error contributed perturbation of the data. In order to be as transparent as possible, this dissertation has included all the sources of error which were observed and which were discernable.

APPENDIX A – TRIGGER WORDS

A.1 CANDIDATE TRIGGER WORDS IDENTIFIED FROM THE LITERATURE

Word	Frequency in Dataset	Prevalence in Messages with Interventions
symptoms	88	0.33
schizophrenia	61	0.44
behavior	9	0.78
thoughts	90	0.4
mental	102	0.32
hard	214	0.31
trouble	36	0.47
thinking	79	0.35
emergency	7	0.57
doctor	104	0.33
voices	202	0.37
normal	35	0.34
negative	38	0.5
disease	25	0.32
different	63	0.43
safety	3	0.67
don't	83	0.39
care	131	0.31
anymore	19	0.47
live	104	0.38
point	69	0.32
teach	6	0.33
depressed	53	0.4
anxious	24	0.42

anxiety	66	0.3
depression	71	0.38
disability	10	0.3
effect	48	0.44
effects	23	0.39
medication	119	0.34
medications	25	0.4
med	362	0.32
meds	106	0.27
harm	17	0.35

A.2 CANDIDATE TRIGGER WORDS IDENTIFIED FROM DATASET

Word	**Frequency In Messages with Interventions**
feel	169
know	155
good	152
time	136
people	117
think	107
help	99
work	94
voices	91
hope	79
everyone	72
try	70
better	70
sometimes	67
hard	67
life	62
job	60
illness	60

APPENDIX B – RECRUITING AND TRAINING DOCUMENTS

B.1 IN-PERSON RECRUITING HANDOUTS

Discussion Forum Moderating Study

The purpose of this research study is to determine whether an automated decision-aiding system can help experts be more efficient when making decisions while moderating controlled access online social network discussion forums.

What you'll do:
Make a series of judgements on whether messages from discussion forum deserve a moderator intervention response or not.

Time Commitment:
30-45 minutes

Location:
Information Science Building #723

Compensation:
$20

Contact:
William Garrard
Email **wcg5003@gmail.com** to schedule a session.

Qualification:
Must be a MSW/Master of Counseling student

Entirely anonymous

B.2 CONSENT SCRIPT

The purpose of this research study is to determine whether an automated decision-aiding system can help experts be more efficient when making decisions while moderating controlled access online social network discussion forums.

For that reason, we will be conducting a study in which participants will be asked to make a series of judgements on whether messages from such a discussion forum deserve a moderator intervention response or not. The experiment should take less than an hour to complete.

If you are willing to participate, we will ask about your familiarity with automated decision aiding systems and your experience in social work. There are no foreseeable risks associated with this study, nor are there any direct benefits to you. Each participant will receive $20 upon completion of the study.

This is an entirely anonymous study. Your responses will not be identifiable in any way. All responses are confidential and will be kept secured. Your participation is voluntary and you may withdraw from this study at any time. No signature is necessary, consent will be obtained verbally.

This study is being conducted by William Garrard, who can be reached at (724) 309-5430 / wcg5003@gmail.com if you have any questions.

B.3 TRAINING FOR CONTROL CASE

Discussion Forum Moderator Training

Situation:

Imagine yourself as the moderator of an online discussion forum used by many people with schizophrenia. Your job is to keep the discussion organized and positive. When you write messages on the board from your position of authority it is called an **intervention**.

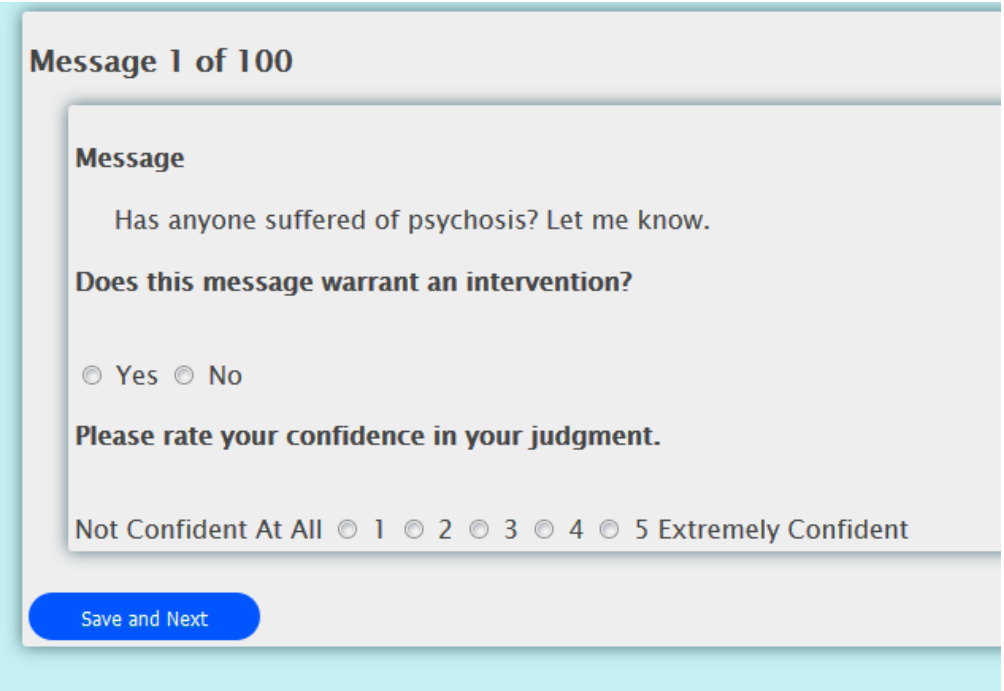
You **want** to make interventions when it is **necessary** because it preserves the organized and positive nature of the board.

You **do not want** to make interventions when it is **not necessary** because it breaks down the organic nature of discussion.

Task:

You are going to be shown 100 messages from the discussion board. The messages are from real people with schizophrenia. For each message, you will decide if it deserves an intervention or not. You will also rate your confidence.

Before and after the messages are shown, there are brief surveys for you to fill out about yourself and your experience in this session.



Discussion Forum Moderator Training

Situation:

Imagine yourself as the moderator of an online discussion forum used by many people with schizophrenia. Your job is to keep the discussion organized and positive. When you write messages on the board from your position of authority it is called an **intervention**.

You **want** to make interventions when it is **necessary** because it preserves the organized and positive nature of the board.

You **do not want** to make interventions when it is **not necessary** because it breaks down the organic nature of discussion.

Task:

You are going to be shown 100 messages from the discussion board. The messages are from real people with schizophrenia. For each message, you will decide if it deserves an intervention or not. You will also rate your confidence.

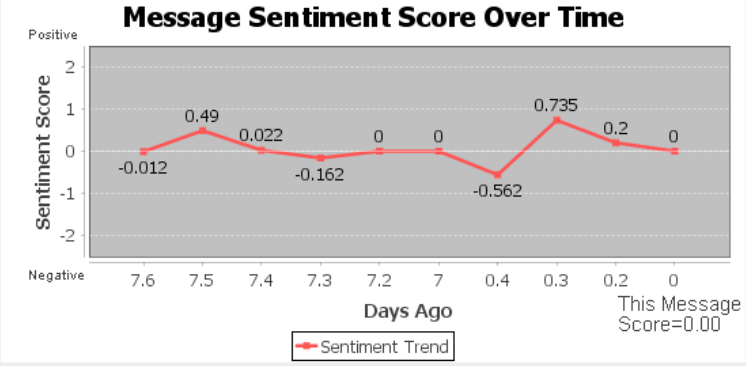
Tools:

For each message, you will see a visualization of the message author's recent past messages. The messages will be represented as points on a graph of the message's **sentiment polarity**. This is a calculated score for how **positive or negative** a message is. A positive score means positive sentiment and a negative score means negative sentiment. The points on the graph will be labeled in **relative days ago** from the date of the current message.

You will also see some **trigger word highlighting** in the messages. These are words identified as potential indicators of intervention necessity. These will be highlighted in **red**.

Before and after the messages are shown, there are brief surveys for you to fill out about yourself and your experience in this session.

Message 1 of 100



Message

How is it hearing voices during prgnancy

Does this message warrant an intervention?

Yes No

Please rate your confidence in your judgment.

Not Confident At All 1 2 3 4 5 Extremely Confident

Save and Next

APPENDIX C - APPARATUS

The experimental system has been developed as a Java Web Application using the NetBeans IDE. It has been deployed on an Apache Tomcat web server and uses a MySQL database for data storage. Client side scripting is handled by JavaScript. Visualizations are generated in real time using JFreeChart. Sentiment score calculation is performed using Semantria for Excel and was precomputed.

REFERENCES

[1] "communication, n." Oxford English Dictionary.

[2] D. M. Boyd and N. B. Ellison, "Social Network Sites: Definition, History, and Scholarship," *J. Comput. Commun.*, vol. 13, no. 1, pp. 210–230, 2007.

[3] "forum, n." Oxford English Dictionary.

[4] J. J. Garrett, "Ajax: A new approach to web applications," pp. 1–5, 2005.

[5] S. Asunka, H. S. Chae, B. Hughes, and G. Natriello, "Understanding academic information seeking habits through analysis of web server log files: the case of the teachers college library website," *J. Acad. Librariansh.*, vol. 35, no. 1, pp. 33–45, 2009.

[6] B. J. Jansen, "Search log analysis: What it is, what's been done, how to do it," *Libr. Inf. Sci. Res.*, vol. 28, no. 3, pp. 407–432, 2006.

[7] A. Phippen, L. Sheppard, S. Furnell, a. Phippen, L. Sheppard, and S. Furnell, "A practical evaluation of Web analytics," *Internet Res.*, vol. 14, no. 4, pp. 284–293, 2004.

[8] Y. Zhang, B. J. Jansen, and A. Spink, "Time series analysis of a Web search engine transaction log," *Inf. Process. Manag.*, vol. 45, no. 2, pp. 230–245, 2009.

[9] S. Park, J. H. Lee, and H. J. Bae, "End user searching: A Web log analysis of NAVER, a Korean Web search engine," *Libr. Inf. Sci. Res.*, vol. 27, no. 2, pp. 203–221, 2005.

[10] F. Schneider, A. Feldmann, B. Krishnamurthy, and W. Willinger, “Understanding online social network usage from a network perspective,” in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, 2009, pp. 35–48.

[11] S. Asunka, H. S. Chae, B. Hughes, and G. Natriello, “Understanding academic information seeking habits through analysis of web server log files: the case of the teachers college library website,” *J. Acad. Librariansh.*, vol. 35, no. 1, pp. 33–45, 2009.

[12] R. Atterer, M. Wnuk, and A. Schmidt, “Knowing the user’s every move: user activity tracking for website usability evaluation and implicit interaction,” in *Proceedings of the 15th international conference on World Wide Web*, 2006, pp. 203–212.

[13] W. Fang, “Using Google Analytics for Improving Library Website Content and Design: A Case Study,” *Libr. Philos. Pract.*, vol. 2007, pp. 1–17, 2007.

[14] R. Atterer and A. Schmidt, “Tracking the interaction of users with AJAX applications for usability testing,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2007, pp. 1347–1350.

[15] E. Kiciman and B. Livshits, “AjaxScope: a platform for remotely monitoring the client-side behavior of Web 2.0 applications,” in *ACM SIGOPS Operating Systems Review*, 2007, vol. 41, no. 6, pp. 17–30.

[16] A. Pak and P. Paroubek, “Twitter as a Corpus for Sentiment Analysis and Opinion Mining.,” *Lrec*, pp. 1320–1326, 2010.

[17] S. Baccianella, A. Esuli, and F. Sebastiani, “SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.,” in *LREC*, 2010, vol. 10, pp. 2200–2204.

[18] B. Pang and L. Lee, “A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts,” in *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, 2004, vol. 2, no. 1–2, pp. 1–135.

[19] T. Wilson, J. Wiebe, P. Hoffmann, and B. Columbia, “Recognizing Contextual Polarity in Phrase-level Sentiment Analysis,” in *Proceedings of the Conference on Human Language*

Technology and Empirical Methods in Natural Language Processing, 2005, no. October, pp. 347–354.

[20] T. Nasukawa, "Sentiment Analysis : Capturing Favorability Using Natural Language Processing Definition of Sentiment Expressions," *2nd Int. Conf. Knowl. capture*, pp. 70–77, 2003.

[21] N. Godbole and M. Srinivasaiah, "Large-scale sentiment analysis for news and blogs," *Conf. Weblogs Soc. Media (ICWSM 2007)*, pp. 219–222, 2007.

[22] H. Kennedy, "Perspectives on Sentiment Analysis.," *J. Broadcast. Electron. Media*, vol. 56, no. 4, pp. 435–450, 2012.

[23] H. Chen, "Sentiment Analysis," in *Dark Web*, vol. 30, Springer New York, 2012, pp. 171–201.

[24] Lexalytics, "Semantria About." [Online]. Available: https://www.lexalytics.com/about.

[25] Lexalytics, "Semantria." [Online]. Available: https://www.lexalytics.com/.

[26] C. M. Kurniawan and D. P. Koesrindartoto, "Sentiment Analysis of Consumer Goods and Mining Sectors Index Performance based on Online Forum and Online News Activities in Indonesia," *Igarss 2014*, no. 1, pp. 1–5, 2014.

[27] H. Pangemanan and D. P. Koesrindartoto, "Indonesia Capital Market Behavior Using Sentiment Measurement in Stockbit Conversation," 2015.

[28] M. Araújo, A. C. M. Pereira, B. Horizonte, J. C. S. Reis, A. C. M. Pereira, and F. Benevenuto, "An Evaluation of Machine Translation for Multilingual Sentence-level Sentiment Analysis," 2016.

[29] J. Peisenieks, R. Skadiņš, and R. Skadicnš, "Uses of Machine Translation in the Sentiment Analysis of Tweets," in *Human Language Technologies-The Baltic Perspective: Proceedings of the Sixth International Conference Baltic HLT 2014*, 2014, vol. 268, p. 126.

[30] E. Fazzion, P. Las Casas, G. Gonc, P. Las Casas, G. Gonçalves, R. Melo-Minardi, and W. Meira Jr, “Open Weekend and Rating Prediction Based on Visualization Techniques,” in *Proc. IEEE Int. Conf. on Visual Analytics Science and Technology (VAST Challenge Paper)*, 2013, pp. 1–2.

[31] K. V. Renaldi and D. P. Koesrindartoto, “Get Some Returns through Online News Sentiment on Big Cap Market,” no. 2007, pp. 160–166, 2015.

[32] D. Wicks, D. Wicks, A. Lumpe, D. Wicks, R. Henrikson, and N. Baliram, “Semantic Text Theme Generation in Collaborative Online Learning Environments,” in *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, 2015, vol. 2015, no. 1, pp. 1837–1842.

[33] D. Georgiou, A. MacFarlane, and T. Russell-rose, “Extracting sentiment from healthcare survey data: An evaluation of sentiment analysis tools,” in *Science and Information Conference (SAI), 2015*, 2015, pp. 352–361.

[34] A. Linus, P. Lawrence, and L. Lawrence, “Reliability of Sentiment Mining Tools: A comparison of Semantria and Social Mention,” pp. 1–13, 2014.

[35] J. M. Van Aggelen and J. M. van Aggelen, “Concurrent Validity and Consistency of Social Media Sentiment Analysis Tools,” pp. 1–8, 2015.

[36] R. M. Marra, R. M. Marra, J. L. Moore, J. L. Moore, A. K. Klimczak, and A. K. Klimczak, “Content analysis of online discussion forums: A comparative analysis of protocols,” *Educ. Technol. Res. Dev.*, vol. 52, no. 2, pp. 23–40, 2004.

[37] H. Kanuka and T. Anderson, “Online social interchange, discord, and knowledge construction,” *Int. J. E-Learning Distance Educ.*, vol. 13, no. 1, pp. 57–74, 2007.

[38] P. B. de Laat, “Navigating between chaos and bureaucracy: Backgrounding trust in open-content communities,” in *Social Informatics*, Springer, 2012, pp. 543–557.

[39] P. B. de Laat, “Coercion or empowerment? Moderation of content in Wikipedia as

‘essentially contested’ bureaucratic rules,” *Ethics Inf. Technol.*, vol. 14, no. 2, pp. 123–135, 2012.

[40] L. Postma and A. S. Blignaut, “Silencing dissent in an online discussion forum of a higher education institution,” *TD J. Transdiscipl. Res. South. Africa*, vol. 9, no. 2, pp. 277–294, 2013.

[41] R. Mason, “Evaluation methodologies for computer conferencing applications,” in *Collaborative learning through computer conferencing*, Springer, 1992, pp. 105–116.

[42] D. R. Garrison, T. Anderson, and W. Archer, “Critical thinking, cognitive presence, and computer conferencing in distance education,” *Am. J. Distance Educ.*, vol. 15, no. 1, pp. 7–23, 2001.

[43] A. Schmidt, “Implicit human computer interaction through context,” *Pers. Technol.*, vol. 4, no. 2–3, pp. 191–199, 2000.

[44] E. K. McCreary, “Three behavioral models for computer-mediated communication,” *Online Educ. Perspect. a new Environ.*, pp. 117–130, 1990.

[45] F. Henri, *Computer conferencing and content analysis*. Springer, 1992.

[46] G. Burnett, “Information exchange in virtual communities: a typology,” *Inf. Res.*, vol. 5, no. 4, 2000.

[47] G. Burnett and H. Buerkle, “Information Exchange in Virtual Communities: A Comparative Study,” *J. Comput. Commun.*, vol. 9, no. 2, p. 0, 2004.

[48] D. R. Garrison, T. Anderson, and W. Archer, “Critical inquiry in a text-based environment: Computer conferencing in higher education,” *internet High. Educ.*, vol. 2, no. 2, pp. 87–105, 1999.

[49] C. N. Gunawardena, C. A. Lowe, T. Anderson, C. H. Sides, D. L. Carson, P. V Anderson, V. a Book, T. C. Dixon, J. S. Harris, F. T. Kiley, T. E. Pearsall, J. C. Redish, J. E. Harmon, and C. J. M. Jansen, “Analysis of a global online debate and the development of

an interaction analysis model for examining social construction of knowledge in computer conferencing," *J. Educ. Comput. Res.*, vol. 17, no. 4, pp. 397–431, 1997.

[50] D. R. Newman, B. Webb, and C. Cochrane, "A content analysis method to measure critical thinking in face-to-face and computer supported group learning," *Interpers. Comput. Technol.*, vol. 3, no. 2, pp. 56–77, 1995.

[51] D. Hume, H. G. Blocker, and C. W. Hendel, *An inquiry concerning human understanding*, vol. 49. Bobbs-Merrill Indianapolis, 1955.

[52] R. B. Adler, G. R. Rodman, and C. Cropley, *Understanding human communication*. Holt, Rinehart, and Winston Fort Worth, 1991.

[53] D. Kantor, *Reading the room: group dynamics for coaches and leaders*, 1st ed. San Francisco: Jossey-Bass, 2012.

[54] J. Prager, B. Pang, and L. Lee, "Opinion Mining and Sentiment Analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 1–2, pp. 1–135, Jan. 2008.

[55] A. S. Brunker, Q. V. Nguyen, A. J. Maeder, R. Tague, G. S. Kolt, T. N. Savage, C. Vandelanotte, M. J. Duncan, C. M. Caperchione, R. R. Rosenkranz, A. Van Itallie, W. K. Mummery, and others, "A Time-based Visualization for Web User Classification in Social Networks," in *Proceedings of the 7th International Symposium on Visual Information Communication and Interaction*, 2014, p. 98.

[56] K. A. Cook and J. J. Thomas, "Illuminating the path: The research and development agenda for visual analytics," 2005.

[57] J. J. Thomas, K. a. Cook, and others, "A visual analytics agenda," *Comput. Graph. Appl. IEEE*, vol. 26, no. 1, pp. 10–13, 2006.

[58] D. a. Keim, F. Mansmann, D. Oelke, and H. Ziegler, "Visual analytics: Combining automated discovery with interactive visualizations," in *Discovery Science*, 2008, vol. 5255 LNAI, pp. 2–14.

[59] T. A. Elena Zudilova-Seinstra, E. Zudilova-Seinstra, T. Adriaansen, and R. van Liere, “Trends in interactive Visualization,” *Trends Interact. Vis. Adv. Inf. Knowl. Process.*, vol. 1, pp. 1–397, 2009.

[60] R. Ball and C. North, “Realizing embodied interaction for visual analytics through large displays,” *Comput. Graph.*, vol. 31, no. 3, pp. 380–400, 2007.

[61] L. C. Freeman, “Visualizing social networks,” *J. Soc. Struct.*, vol. 1, no. 1, p. 4, 2000.

[62] N. Henry, J.-D. Fekete, and M. J. McGuffin, “NodeTrix: a hybrid visualization of social networks,” *Vis. Comput. Graph. IEEE Trans.*, vol. 13, no. 6, pp. 1302–1309, 2007.

[63] J. Heer and D. Boyd, “Vizster: Visualizing online social networks,” in *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, 2005, pp. 32–39.

[64] R. Tague, A. Maeder, and Q. V. Nguyen, “Interactive visualisation of time-based vital signs,” in *Advances in Visual Computing*, Springer, 2010, pp. 545–553.

[65] W. Aigner, S. Miksch, W. Müller, H. Schumann, and C. Tominski, “Visualizing time-oriented data-a systematic view,” *Comput. Graph.*, vol. 31, no. 3, pp. 401–409, 2007.

[66] B. Shneiderman, “The eyes have it: A task by data type taxonomy for information visualizations,” in *Visual Languages, 1996. Proceedings., IEEE Symposium on*, 1996, pp. 336–343.

[67] C. Tominski, W. Aigner, S. Miksch, W. Muller, H. Schumann, and C. Tominski, “Visual methods for analyzing time-oriented data,” *Vis. Comput. Graph. IEEE Trans.*, vol. 14, no. 1, pp. 47–60, 2008.

[68] W. Aigner, S. Miksch, H. Schumann, and C. Tominski, *Visualization of Time-Oriented Data*. 2011.

[69] J. Lin, E. Keogh, L. Wei, and S. Lonardi, “Experiencing SAX: a novel symbolic representation of time series,” *Data Min. Knowl. Discov.*, vol. 15, no. 2, pp. 107–144, 2007.

[70] J. Lin, E. Keogh, S. Lonardi, J. P. Lankford, and D. M. Nystrom, “Visually mining and monitoring massive time series,” in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, vol. 1, pp. 460–469.

[71] Y. Seong and A. M. Bisantz, “Assessment of operator trust in and utilization of automated decision aids under different framing conditions,” *Proc. Hum. Factors Ergon. Soc. ...Annual Meet.*, vol. 1, p. 5, 2000.

[72] Y. Seong, A. M. Bisantz, and A. M. Bisantz, “Judgment and Trust in Conjunction with Automated Decision Aids: A Theoretical Model and Empirical Investigation,” *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 46, no. 3, pp. 423–427, 2002.

[73] P. Madhavan and D. A. Wiegmann, “Expertise Levels of Human versus Automated Decision Aids Influence Response Biases in a Visual Search Task,” *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 50, no. 3, pp. 230–234, 2006.

[74] J. E. Bahner, A.-D. Hüper, D. Manzey, J. E. B. Ã, A. Hu, and D. Manzey, “Misuse of automated decision aids: Complacency, automation bias and the impact of training experience,” *Int. J. Hum. Comput. Stud.*, vol. 66, no. 9, pp. 688–699, 2008.

[75] K. L. Mosier and L. J. Skitka, “Human decision makers and automated decision aids: Made for each other,” *Autom. Hum. Perform. Theory Appl.*, pp. 201–220, 1996.

[76] D. Manzey, J. Reichenbach, and L. Onnasch, “Human performance consequences of automated decision aids: The impact of degree of automation and system experience,” *J. Cogn. Eng. Decis. Mak.*, p. 1555343411433844, 2012.

[77] K. Van Dongen and P. P. Van Maanen, “A framework for explaining reliance on decision aids,” *Int. J. Hum. Comput. Stud.*, vol. 71, no. 4, pp. 410–424, 2013.

[78] R. A. Miller, H. E. Pople Jr, and J. D. Myers, “Internist-I, an experimental computer-based diagnostic consultant for general internal medicine,” *N. Engl. J. Med.*, vol. 307, no. 8, pp. 468–476, 1982.

[79] D. A. Ferrucci, “Introduction to This is Watson,” *IBM J. Res. Dev.*, vol. 56, no. 3.4, p. 1:1-1:15, May 2012.

[80] N. Leske, “Doctors Seek Help on Cancer Treatment from IBM Supercomputer,” *Reuters*, 2013. [Online]. Available: http://in.reuters.com/article/ibm-watson-cancer-idINDEE9170G120130208. [Accessed: 02-Feb-2016].

[81] J. N. John Natale, Christine Douglass, “MD Anderson Taps IBM Watson to Power ‘Moon Shots’ Mission Aimed at Ending Cancer, Starting with Leukemia,” *IBM News Room*, 2013. [Online]. Available: http://www-03.ibm.com/press/us/en/pressrelease/42214.wss. [Accessed: 02-Feb-2016].

[82] M. Devarakonda, D. Zhang, C. H. Tsou, and M. Bornea, “Problem-oriented patient record summary: An early report on a Watson application,” *2014 IEEE 16th Int. Conf. e-Health Networking, Appl. Serv. Heal. 2014*, pp. 281–286, 2015.

[83] P. Ruchikachorn, J. J. Liang, M. Devarakonda, and K. Mueller, “Watson -Aided Non-Linear Problem-Oriented Clinical Visit Preparation on Tablet Computer,” *Vizualizing Electron. Heal. Rec. Data*, pp. 1–4, 2014.

[84] JD, “How Experts Make Decisions,” *Sources of Insight*, 2007. [Online]. Available: http://sourcesofinsight.com/how-experts-make-decisions/.

[85] F. T. Penney, “How Do Experts Make Decisions,” *MUSC*, 2013.

[86] G. Klein, *Sources of power: How people make decisions*. MIT press, 1999.

[87] H. E. Montgomery, R. E. Lipshitz, and B. E. Brehmer, *How professionals make decisions*. Lawrence Erlbaum Associates Publishers, 2005.

[88] M. Cox, D. M. Irby, and J. L. Bowen, “Educational strategies to promote clinical diagnostic reasoning,” *N. Engl. J. Med.*, vol. 355, no. 21, pp. 2217–2225, 2006.

[89] M. Groves, P. O’Rourke, and H. Alexander, “The clinical reasoning characteristics of diagnostic experts,” *Med. Teach.*, vol. 25, no. 3, pp. 308–313, 2003.

[90] A. Dijksterhuis, M. W. Bos, L. F. Nordgren, and R. B. van Baaren, “On Making the Right Choice: The Deliberation-Without-Attention Effect,” *Science (80-.).*, vol. 311, no. 5763, pp. 1005–1007, 2006.

[91] R. Williams, “How Can We Make Better Decisions?,” *Psychology Today*, 2011. [Online]. Available: https://www.psychologytoday.com/blog/wired-success/201109/how-can-we-make-better-decisions.

[92] J. Shanteau, “How much information does an expert use? Is it relevant?,” *Acta Psychol. (Amst).*, vol. 81, no. 1, pp. 75–86, 1992.

[93] J. Shanteau, “Psychological characteristics and strategies of expert decision makers,” *Acta Psychol. (Amst).*, vol. 68, no. 1–3, pp. 203–215, 1988.

[94] A. S. Elstein, A. Schwartz, and A. Schwarz, “Clinical problem solving and diagnostic decision making: selective review of the cognitive literature.,” *Br. Med. J.*, vol. 324, no. March, pp. 729–732, 2002.

[95] E. Fargiorgio, “Sentiment analysis explained1.” 2016.

[96] M. Obretenov, “Sentiment analysis explained2.” 2016.

[97] T. Winograd, “Understanding natural language,” *Cogn. Psychol.*, vol. 3, no. 1, pp. 1–191, 1972.

[98] C. Lamorte, “Personal Letter on Trigger Words.” 2016.

[99] B. Neely, “Personal Letter on Trigger Words.” 2016.

[100] Mayo_Clinic_Staff, “Schizophrenia Symptoms,” 2014. [Online]. Available: http://www.mayoclinic.org/diseases-conditions/schizophrenia/basics/symptoms/con-20021077.

[101] J. Goldberg, “Schizophrenia Symptoms,” 2015. [Online]. Available: http://www.webmd.com/schizophrenia/guide/schizophrenia-symptoms.

[102] B. Neely, “Trigger Words Review.” 2016.

[103] C. Lamorte, “Trigger Words Review.” 2016.

[104] M. Rauktis, “Letter from Mary Rauktis.” 2016.

[105] A. Rotondi, “Health Care Innovation Challenge: Improving Quality and Reducing Cost in Schizophrenia Care with New Technologies and New Personnel.” Feinstein Institute for Medical Research: Centers for Medicare & Medicaid Services, 2015.

[106] D. Armstrong, A. Gosling, J. Weinman, and T. Marteau, “The Place of Inter-Rater Reliability in Qualitative Research: An Empirical Study,” *Sociology*, vol. 31, no. 3, pp. 597–606, 1997.

[107] K. L. Gwet, *Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters*, Fourth. Gaithersburg, MD: Advanced Analytics, LLC, 2014.

[108] J. Cohen, “A coefficient of agreement for nominal scales,” *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

[109] C. G. Blanchard, M. S. Labrecque, J. C. Ruckdeschel, and E. B. Blanchard, “Information and decision-making preferences of hospitalized adult cancer patients,” *Soc. Sci. Med.*, vol. 27, no. 11, pp. 1139–1145, 1988.

[110] S. Eckermann and A. R. Willan, “Expected value of information and decision making in HTA,” *Health Econ.*, vol. 16, no. 2, pp. 195–209, 2007.

[111] J. G. March, “Ambiguity and accounting: The elusive link between information and decision making,” *Accounting, Organ. Soc.*, vol. 12, no. 2, pp. 153–168, 1987.

[112] Q. Gong, Y. Yang, and S. Wang, “Information and decision-making delays in MRP, KANBAN, and {CONWIP},” *Int. J. Prod. Econ.*, vol. 156, pp. 208–213, 2014.

[113] P. Todd and I. Benbasat, “The Use of Information in Decision Making: An Experimental Investigation of the Impact of Computer-Based Decision Aids,” *MIS Q.*, vol. 16, no. 3, pp. 373–393, 1992.